PCT/NL 2004 / 0 0 0 8 0 5

Europäisches Patentamt | European Patent Office | Office européen des brevets

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. | Patent application No. | Demande de brevet n°

03078613.1

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 03078613.1
Demande no:

Anmeldetag:
Date of filing: 18.11.03
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Vironovative B.V.
Burgemeester Oudlaan 50
3062 PA Rotterdam
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Novel atypical pneumonia-causing virus

In Anspruch genommene Prioriät(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

C12N7/00

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL
PT RO SE SI SK TR LI

Title: Novel atypical pneumonia-causing virus

The invention relates to the field of virology.

The SARS outbreak of 2002-2003 has prompted a search for related viruses that may have previously caused atypical pneumonias or that may do so in the future. A respiratory illness (atypical pneumonia) was diagnosed in an 8 months old patient that could not be attributed to SARS (Severe Acute Respiratory Syndrome) virus or any othe known viral infection. The patient tested negative for influenza, parainfluenza, mumps and RSV and yet the disease was identified to be caused by a virus which closely resembled SARS.

For being able to trace its origin, monitor its epidemiology and prevent possible spreading of the disease, it is of great importance to be able to recognise viral causes of pneumonia in an early stage. Especially, if severe diseases are found to be caused by viruses, it is necessary to detect the identity of the virus as soon as possible, in order to develop diagnostic tools and possibly therapies. The SARS epidemy has shown that it is paramount for prevention of spread of the disease to be able to get an early diagnosis in order to take timely and effective isolation measures and initiate quarantine precautions. Only then, world-wide contaminations can be prevented.

Furthermore, identification of the viral cause for the disease enables development of vaccines, which can be used prophylactically to protect people who are a risk of being infected. And, finally, knowledge of the viral cause enables to develop therapeutic measures.

Thus, there is great need in developing diagnostic tools and therapies for viral pneumonias in general, and particular to a novel disease-causing infectious agent, especially when this agent appears to be a virus.

The invention provides the nucleotide sequence of an isolated essentially mammalian positive-sense single stranded RNA virus belonging to the Coronaviruses, which is the causative factor for the new disease, hereinafter referred to as EMCR-CoV and the disease being referred to as EMCR-CoV-caused pneumonia. A virus according tc

the invention is isolatable from a human with respiratory tract disease such as, but not limited to, atypical pneumonia.

From a phylogenetic analysis of the Matrix and Nucleocapsid gene sequences of the virus (Fig. 1a and 1b) it appears that the virus is a distinct member of the group formed by PEDV (porcine epidemic diarrhea virus), HCoV-229E (human coronavirus 229E), PRCoV (porcine respiratory coronavirus), TGEV (transmissible gastroenteritis virus), CaCoV (Canine coronavirus) and FeCoV (feline coronavirus). In general, human coronavirus 229E seems to be the closest relative (at least for the Matrix and Nucleocapsid proteins).

Although phylogenetic analyses provide a convenient method of identifying a virus, several other possibly more straightforward albeit somewhat more coarse methods for identifying said virus or viral proteins or nucleic acids from said virus are herein also provided. As a rule of thumb an EMCR-Coronavirus can be identified by the percentages of homology of the virus, proteins or nucleic acids to be identified in comparison with viral proteins or nucleic acids identified herein by sequence. It is generally known that virus species, especially RNA virus species, often constitute a quasi species wherein a cluster of said viruses displays heterogeneity among its members. Thus it is expected that each isolate may have a somewhat different percentage relationship with the sequences of the isolate as provided herein.

When one wishes to compare a virus isolate with the sequences as listed in figure 3, the invention provides an isolated essentially mammalian positive-sense single stranded RNA virus (EMCR-CoV) belonging to the Coronaviruses and identifiable as phylogenetically corresponding thereto by determining a nucleic acid sequence of said virus and determining that said nucleic acid sequence has a percentage nucleic acid identity to the sequences as listed higher than the percentages identified herein for the nucleic acids as identified herein below in comparison with PEDV, 229E, PRCoV, TGEV CaCoV and FeCoV. Likewise, an isolated essentially mammalian positive-sense single stranded RNA virus (EMCR-CoV) belonging to the Coronaviruses and identifiable as phylogenetically corresponding thereto by determining an amino acid sequence of said virus and determining that said amino acid sequence has a percentage amino acid homology to the sequences as listed which is essentially higher than the percentages provided herein in comparison with PEDV, 229E, PRCoV, TGEV, CaCoV and FeCoV.

With the provision of the sequence information of this EMCR-Coronavirus (EMCR-CoV), the invention provides diagnostic means and methods, prophylactic mean

and methods and therapeutic means and methods to be employed in the diagnosis, prevention and/or treatment of disease, in particular of respiratory disease (atypical pneumonia), in particular of mammals, more in particular in humans associated with infection by this virus. In virology, it is most advisory that diagnosis, prophylaxis and/o treatment of a specific viral infection is performed with reagents that are most specific for said specific virus causing said infection. In this case this means that it is preferred that said diagnosis, prophylaxis and/or treatment of an EMCR-CoV virus infection is performed with reagents that are most specific for EMCR-CoV virus. This by no means however excludes the possibility that less specific, but sufficiently cross-reactive reagents are used instead, for example because they are more easily available and sufficiently address the task at hand.

The invention for example provides a method for virologically diagnosing an EMCR-CoV infection of an animal, in particular of a mammal, more in particular of a human being, comprising determining in a sample of said animal the presence of a viral isolate or component thereof by reacting said sample with an EMCR-CoV specific nuclei acid or antibody according to the invention, and a method for serologically diagnosing a EMCR-CoV infection of a mammal comprising determining in a sample of said mammal the presence of an antibody specifically directed against an EMCR-CoV virus or component thereof by reacting said sample with an EMCR-CoV virus-specific proteinaceous molecule or fragment thereof or an antigen according to the invention.

The invention also provides a diagnostic kit for diagnosing an EMCR-CoV infection comprising an EMCR-CoV virus, an EMCR-CoV virus-specific nucleic acid, proteinaceous molecule or fragment thereof, antigen and/or an antibody according to the invention, and preferably a means for detecting said EMCR-CoV virus, EMCR-CoV virus-specific nucleic acid, proteinaceous molecule or fragment thereof, antigen and/or an antibody, said means for example comprising an excitable group such as a fluorophore or enzymatic detection system used in the art (examples of suitable diagnostic kit format comprise IF, ELISA, neutralization assay, RT-PCR assay). To determine whether an as yet unidentified virus component or synthetic analogue thereo such as nucleic acid, proteinaceous molecule or fragment thereof can be identified as EMCR-CoV-virus-specific, it suffices to analyse the nucleic acid or amino acid sequence of said component, for example for a stretch of said nucleic acid or amino acid, preferably of at least 10, more preferably at least 25, more preferably at least 40 nucleotides or amino acids (respectively), by sequence homology comparison with the

provided EMCR-CoV viral sequences and with known non-EMCR-CoV viral sequences (human coronavirus 299E is preferably used) using for example phylogenetic analyses as provided herein. Depending on the degree of relationship with said EMCR-CoV or non-EMCR-CoV viral sequences, the component or synthetic analogue can be identified.

The invention thus provides the nucleotide sequence of a novel etiological agent, an isolated essentially mammalian positive-sense single stranded RNA virus (herein also called EMCR-CoV virus) belonging to the Coronaviridae family, and EMCR-CoV virus-specific components or synthetic analogues thereof.

Coronaviruses were first isolated from chickens in 1937, while the first human coronavirus was propagated *in vitro* by Tyrell and Bonoe in 1965. There are now about 13 species in this family, which infect cattle, pigs, rodents, cats, dogs, birds and man. Coronavirus particles are irregularly shaped, about 60-220 nm in diameter, with an outer envelope bearing distinctive, 'club-shaped' peplomers (about 20 nm long and 10 nm wide at the distal end). This 'crown-like' appearance give the family its name. The envelope carries two glycoproteins: S, the spike glycoprotein which is involved in cell fusion and is a major antigen, and M, the membrane glycoprotein, which is involved in budding and envelope formation. The genome is associated with a basic phosphoprotein, designated N. The genome of coronaviruses, a single stranded positive-sense RNA strand, is typically 27-31 Kb long and contains a 5' methylated cap and a 3' poly-A tail, by which it can directly function as an mRNA in the infected cell. Initially the 5' ORF 1 (about 20 Kb) is translated to produce a viral polymerase, which then produces a full length negative sense strand. This is used as a template to produce mRNA as a 'nested set' of transcripts, all with identical 5' non-translated leader sequence of 72 nucleotides and coincident 3' polyadenylated ends. Each mRNA thus produced is monocistronic, the genes at the 5' end being translated from the longest mRNA and so on. These unusual cytoplasmic structures are produced not by splicing, but by the polymerase during transcription. Between each of the genes there is a repeated intergenic sequence – AACUAAAC – which interacts with the transcriptase plus cellular factors to splice the leader sequence onto the start of each ORF. In some coronaviruses there are about 8 ORFs, coding for the proteins mentioned above, but also for a heamagglutenin esterase (HE), and several other non-structural proteins.

Newly isolated viruses are phylogenetically corresponding to and thus taxonomically corresponding to EMCR-CoV virus when comprising a gene order and/or amino acid sequence and/or nucleotide sequence sufficiently similar to our prototypic

EMCR-CoV virus. The highest amino acid sequence homology, between EMCR-CoV virus and any of the known other viruses of the same family to date (human coronaviru 299E or Porcine Epidemic Diarrhea Virus) is for parts of the replicase polyprotein 1ab 80-83% (see, for example Fig. 3 sequences D and E; the % homology, and the virus to which the homology is found depend on the region of the replicase that is examined), a can be deduced when comparing the sequences given in figure 3 with sequences of othe viruses, in particular of human coronavirus 299E. Individual proteins or whole virus isolates with, respectively, higher homology than these mentioned maximum values ar considered phylogenetically corresponding and thus taxonomically corresponding to EMCR-CoV virus, and generally will be encoded by a nucleic acid sequence structurall corresponding with a sequence as shown in figure 3. Herewith the invention provides a virus phylogenetically corresponding to the isolated virus of which the sequences are depicted in figure 3.

It should be noted that, similar to other viruses, a certain degree of variation ca be expected to be found between EMCR-CoV-viruses isolated from different sources.

Also, the viral sequence of the EMCR-CoV virus or an isolated EMCR-CoV virus gene as provided herein for example shows less than 95%, preferably less than 90%, more preferably less than 80%, more preferably less than 70% and most preferably less than 65% nucleotide sequence homology or less than 95%, preferably less than 90%, more preferably less than 80%, more preferably less than 70% and most preferably less than 65% amino acid sequence homology with the respective nucleotide or amino acid sequence of the human coronavirus 299E or Porcine Epidemic Diarrhea Virus as for example can be found in Genbank (for example in accession number af304460 (HCoV-299E) or af353511 (PEDV).

Sequence divergence of EMCR-CoV strains around the world may be somewhat higher, in analogy with other coronaviruses.

The term "nucleotide sequence homology" as used herein denotes the presence of homology between two (poly)nucleotides. Polynucleotides have "homologous" sequences if the sequence of nucleotides in the two sequences is the same when aligned for maximum correspondence. Sequence comparison between two or more polynucleotides is generally performed by comparing portions of the two sequences over a comparison window to identify and compare local regions of sequence similarity. The comparison window is generally from about 20 to 200 contiguous nucleotides. The "percentage of sequence homology" for polynucleotides, such as 50, 60, 70, 80, 90, 95, 98, 99 or 100

percent sequence homology may be determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide sequence in the comparison window may include additions or deletions (i.e. gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by: (a) determining the number of positions at which the identical nucleic acid base occurs in both sequences to yield the number of matched positions; (b) dividing the number of matched positions by the total number of positions in the window of comparison; and (c) multiplying the result by 100 to yield the percentage of sequence homology. Optimal alignment of sequences for comparison may be conducted by computerized implementations of known algorithms, or by inspection. Readily available sequence comparison and multiple sequence alignment algorithms are, respectively, the Basic Local Alignment Search Tool (BLAST) (Altschul, S.F. et al. 1990. J. Mol. Biol. 215:403; Altschul, S.F. et al. 1997. Nucleic Acid Res. 25:3389-3402) and ClustalW programs both available on the internet. Other suitable programs include GAP, BESTFIT and FASTA in the Wisconsin Genetics Software Package (Genetics Computer Group (GCG), Madison, WI, USA).

As used herein, "substantially complementary" means that two nucleic acid sequences have at least about 65%, preferably about 70%, more preferably about 80%, even more preferably 90%, and most preferably about 98%, sequence complementarity to each other. This means that the primers and probes must exhibit sufficient complementarity to their template and target nucleic acid, respectively, to hybridise under stringent conditions. Therefore, the primer sequences as disclosed in this specification need not reflect the exact sequence of the binding region on the template and degenerate primers can be used. A substantially complementary primer sequence is one that has sufficient sequence complementarity to the amplification template to result in primer binding and second-strand synthesis.

The term "hybrid" refers to a double-stranded nucleic acid molecule, or duplex, formed by hydrogen bonding between complementary nucleotides. The terms "hybridise" or "anneal" refer to the process by which single strands of nucleic acid sequences form double-helical segments through hydrogen bonding between complementary nucleotides.

The term "oligonucleotide" refers to a short sequence of nucleotide monomers (usually 6 to 100 nucleotides) joined by phosphorous linkages (e.g., phosphodiester, alkyl and aryl-phosphate, phosphorothioate), or non-phosphorous linkages (e.g., peptide,

sulfamate and others). An oligonucleotide may contain modified nucleotides having modified bases (e.g., 5-methyl cytosine) and modified sugar groups (e.g., 2'-O-methyl ribosyl, 2'-O-methoxyethyl ribosyl, 2'-fluoro ribosyl, 2'-amino ribosyl, and the like). Oligonucleotides may be naturally-occurring or synthetic molecules of double- and single-stranded DNA and double- and single-stranded RNA with circular, branched or linear shapes and optionally including domains capable of forming stable secondary structures (e.g., stem-and-loop and loop-stem-loop structures).

The term "primer" as used herein refers to an oligonucleotide which is capable o annealing to the amplification target allowing a DNA polymerase to attach thereby serving as a point of initiation of DNA synthesis when placed under conditions in whicl synthesis of primer extension product which is complementary to a nucleic acid strand is induced, i.e., in the presence of nucleotides and an agent for polymerization such as DNA polymerase and at a suitable temperature and pH. The (amplification) primer is preferably single stranded for maximum efficiency in amplification. Preferably, the primer is an oligodeoxy ribonucleotide. The primer must be sufficiently long to prime tł synthesis of extension products in the presence of the agent for polymerization. The exact lengths of the primers will depend on many factors, including temperature and source of primer. A "pair of bi-directional primers" as used herein refers to one forward and one reverse primer as commonly used in the art of DNA amplification such as in PCR amplification.

The term "probe" refers to a single-stranded oligonucleotide sequence that will recognize and form a hydrogen-bonded duplex with a complementary sequence in a target nucleic acid sequence analyte or its cDNA derivative.

The terms "stringency" or "stringent hybridization conditions" refer to hybridization conditions that affect the stability of hybrids, e.g., temperature, salt concentration, pH, formamide concentration and the like. These conditions are empirically optimised to maximize specific binding and minimize non-specific binding o1 primer or probe to its target nucleic acid sequence. The terms as used include reference to conditions under which a probe or primer will hybridise to its target sequence, to a detectably greater degree than other sequences (e.g. at least 2-fold over background). Stringent conditions are sequence dependent and will be different in different circumstances. Longer sequences hybridise specifically at higher temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm

is the temperature (under defined ionic strength and pH) at which 50% of a complementary target sequence hybridises to a perfectly matched probe or primer. Typically, stringent conditions will be those in which the salt concentration is less than about 1.0 M Na+ ion, typically about 0.01 to 1.0 M Na+ ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes or primers (e.g. 10 to 50 nucleotides) and at least about 60°C for long probes or primers (e.g. greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. Exemplary low stringent conditions or "conditions of reduced stringency" include hybridization with a buffer solution of 30% formamide, 1 M NaCl, 1% SDS at 37°C and a wash in 2x SSC at 40°C. Exemplary high stringency conditions include hybridization in 50% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.1x SSC at 60°C. Hybridization procedures are well known in the art and are described in e.g. Ausubel et al, Current Protocols in Molecular Biology, John Wiley & Sons Inc., 1994.

The term "antibody" includes reference to antigen binding forms of antibodies (e. g., Fab, F (ab) 2). The term "antibody" frequently refers to a polypeptide substantially encoded by an immunoglobulin gene or immunoglobulin genes, or fragments thereof which specifically bind and recognize an analyte (antigen). However, while various antibody fragments can be defined in terms of the digestion of an intact antibody, one of skill will appreciate that such fragments may be synthesized de novo either chemically or by utilizing recombinant DNA methodology. Thus, the term antibody, as used herein, also includes antibody fragments such as single chain Fv, chimeric antibodies (i. e., comprising constant and variable regions from different species), humanized antibodies (i. e., comprising a complementarity determining region (CDR) from a non-human source) and heteroconjugate antibodies (e. g., bispecific antibodies).

In short, the invention provides an isolated essentially mammalian positive-sense single stranded RNA virus (EMCR-CoV) belonging to the Coronaviruses and identifiable as phylogenetically corresponding thereto by determining a nucleic acid sequence of a suitable fragment of the genome of said virus and testing it in phylogenetic tree analyses wherein maximum likelihood trees are generated using 100 bootstraps and 3 jumbles and finding it to be more closely phylogenetically corresponding to a virus isolate having the sequences as depicted in figure 3 than it is corresponding to a virus isolate of PEDV (porcine epidemic diarrhea virus), HCoV-229E (human coronavirus 229E), PRCoV (porcine respiratory coronavirus), TGEV

(transmissible gastroenteritis virus), CaCoV (Canine coronavirus) and FeCoV (feline coronavirus).

Suitable nucleic acid genome fragments each useful for such phylogenetic tree analyses are for example any of the fragments encoding the Matrix protein or the Nucleocapsid protein as disclosed in figure 3, leading to the phylogenetic tree analysis as disclosed herein in figure 1a or 1b.

A suitable open reading frame (ORF) comprises the ORF encoding the viral replicase (ORF 1a). When an overall amino acid identity of at least 60%, preferably of a least 70%, more preferably of at least 80%, more preferably of at least 90%, most preferably of at least 95% of the analysed replicase with the replicase having a sequenc comprising the amino acid fragments A, B, C, D, E, and/or F of figure 3 is found, the analysed virus isolate comprises an EMCR-CoV virus isolate according to the invention

Another suitable open reading frame (ORF) useful in phylogenetic analyses comprises the ORF encoding the Nucleocapsid protein. When an overall amino acid identity of at least 60%, more preferably of at least 70%, more preferably of at least 80% more preferably of at least 90%, most preferably of at least 95% of the analysed Nucleocapsid protein with the Nucleocapsid protein encoded by a sequence comprising (part of) the sequence F of figure 3 is found, the analysed virus isolate comprises an EMCR-CoV isolate according to the invention.

Another suitable open reading frame (ORF) useful in phylogenetic analyses comprises the ORF encoding the Matrix protein. When an overall amino acid identity of at least 60%, more preferably of at least 70%, more preferably of at least 80%, more preferably of at least 90%, most preferably of at least 95% of the analysed Matrix protein with the Matrix protein encoded by a sequence comprising (part of) the sequenc F of figure 3 is found, the analysed virus isolate comprises an EMCR-CoV isolate according to the invention.

Another suitable open reading frame (ORF) useful in phylogenetic analyses comprises the ORF encoding the spike protein S. When an overall amino acid identity of at least 60%, more preferably of at least 70%, more preferably of at least 80%, more preferably of at least 90%, most preferably of at least 95% of the analysed S-protein encoded by a sequence comprising the sequence of translation 2 of E and translation 1 of the F sequence of the S-protein as depicted in figure 3 is found, the analysed virus isolate comprises an EMCR-CoV virus isolate according to the invention. The S ORF of the EMCR-CoV virus seems to be located adjacent to the ORF 1ab (coding for the viral

replicase), which would discriminate an EMCR-CoV viruses from the bovine coronavirus and the murine hepatitis virus, which have a so-called 2a gene and an HE-gene between the S protein and the viral polymerase.

The invention provides among others an isolated or recombinant nucleic acid or virus-specific functional fragment thereof obtainable from a virus according to the invention. The isolated or recombinant nucleic acids comprises the sequences as given in figure 3 or sequences of homologues which are able to hybridise with those under stringent conditions. In particular, the invention provides primers and/or probes suitable for identifying an EMCR-CoV virus nucleic acid.

Furthermore, the invention provides a vector comprising a nucleic acid according to the invention. To begin with, vectors such as plasmid vectors containing (parts of) the genome of the EMCR-CoV virus, virus vectors containing (parts of) the genome of the EMCR-CoV (for example, but not limited thereto, vaccinia virus, retroviruses, baculovirus), or EMCR-CoV virus containing (parts of) the genome of other viruses or other pathogens are provided.

Also, the invention provides a host cell comprising a nucleic acid or a vector according to the invention. Plasmid or viral vectors containing the replicase components of EMCR-CoV virus are generated in prokaryotic cells for the expression of the components in relevant cell types (bacteria, insect cells, eukaryotic cells). Plasmid or viral vectors containing full-length or partial copies of the EMCR-CoV virus genome will be generated in prokaryotic cells for the expression of viral nucleic acids *in-vitro* or *in-vivo*. The latter vectors may contain other viral sequences for the generation of chimeric viruses or chimeric virus proteins, may lack parts of the viral genome for the generation of replication defective virus, and may contain mutations, deletions or insertions for the generation of attenuated viruses.

Infectious copies of EMCR-CoV virus (being wild type, attenuated, replication-defective or chimeric) can be produced upon co-expression of the polymerase components according to the state-of-the-art technologies described above.

In addition, eukaryotic cells, transiently or stably expressing one or more full-length or partial EMCR-CoV virus proteins can be used. Such cells can be made by transfection (proteins or nucleic acid vectors), infection (viral vectors) or transduction (viral vectors) and may be useful for complementation of mentioned wild type, attenuated, replication-defective or chimeric viruses.

A chimeric virus may be of particular use for the generation of recombinant vaccines protecting against two or more viruses. For example, it can be envisaged that EMCR-CoV virus vector expressing one or more proteins of a human metapneumovirus or a human metapneumovirus vector expressing one or more proteins of EMCR-CoV virus will protect individuals vaccinated with such vector against both virus infections. Such a specific chimeric virus is particularly useful in the invention because it is suspected that co-infection of, for instance, human metapneumovirus frequently occurs in coronavirus infected patients. Attenuated and replication-defective viruses may be of use for vaccination purposes with live vaccines as has been suggested for other viruses.

In a preferred embodiment, the invention provides a proteinaceous molecule or coronavirus-specific viral protein or functional fragment thereof encoded by a nucleic acid according to the invention. Useful proteinaceous molecules are for example derived from any of the genes or genomic fragments derivable from a virus according to the invention. Such molecules, or antigenic fragments thereof, as provided herein, are for example useful in diagnostic methods or kits and in pharmaceutical compositions such as sub-unit vaccines and inhibitory peptides. Particularly useful are the viral replicase protein, the spike protein, the matrix protein, the nucleocapsid or antigenic fragments thereof for inclusion as antigen or subunit immunogen, but inactivated whole virus can also be used. Particulary useful are also those proteinaceous substances that are encoded by recombinant nucleic acid fragments that are identified for phylogenetic analyses, of course preferred are those that are within the preferred bounds and metes of ORFs useful in phylogenetic analyses, in particular for eliciting EMCR-CoV virus specific antibodies, whether in vivo (e.g. for protective puposes or for providing diagnostic antibodies) or in vitro (e.g. by phage display technology or another technique useful for generating synthetic antibodies).

Also provided herein are antibodies, be it natural polyclonal or monoclonal, or synthetic (e.g. (phage) library-derived binding molecules) antibodies that specifically react with an antigen comprising a proteinaceous molecule or EMCR-CoV virus-specific functional fragment thereof according to the invention. Such antibodies are useful in a method for identifying a viral isolate as an EMCR-CoV virus comprising reacting said viral isolate or a component thereof with an antibody as provided herein. This can for example be achieved by using purified or non-purified EMCR-CoV virus or parts thereof (proteins, peptides) using ELISA, RIA, FACS or similar formats of antigen detection assays (Current Protocols in Immunology). Alternatively, infected cells or cell cultures

may be used to identify viral antigens using classical immunofluorescence or immunohistochemical techniques. Specifically useful in this respect are antibodies raised against EMCR-CoV virus proteins which are encoded by a nucleotide sequence comprising one or more of the fragments disclosed in figure 3.

Other methods for identifying a viral isolate as an EMCR-CoV virus comprise reacting said viral isolate or a component thereof with a virus specific nucleic acid according to the invention.

In this way the invention provides a viral isolate identifiable with a method according to the invention as a mammalian virus taxonomically corresponding to a positive-sense single stranded RNA virus identifiable as likely belonging to the EMCR-CoV virus genus within the family of Coronaviruses.

The method is useful in a method for virologically diagnosing an EMCR-CoV virus infection of a mammal, said method for example comprising determining in a sample of said mammal the presence of a viral isolate or component thereof by reacting said sample with a nucleic acid or an antibody according to the invention.

Methods of the invention can in principle be performed by using any nucleic acid amplification method, such as the Polymerase Chain Reaction (PCR; Mullis 1987, U.S. Pat. No. 4,683,195, 4,683,202, en 4,800,159) or by using amplification reactions such as Ligase Chain Reaction (LCR; Barany 1991, Proc. Natl. Acad. Sci. USA 88:189-193; EP Appl. No., 320,308), Self-Sustained Sequence Replication (3SR; Guatelli et al., 1990, Proc. Natl. Acad. Sci. USA 87:1874-1878), Strand Displacement Amplification (SDA; U.S. Pat. Nos. 5,270,184, en 5,455,166), Transcriptional Amplification System (TAS; Kwoh et al., Proc. Natl. Acad. Sci. USA 86:1173-1177), Q-Beta Replicase (Lizardi et al., 1988, Bio/Technology 6:1197), Rolling Circle Amplification (RCA; U.S. Pat. No. 5,871,921), Nucleic Acid Sequence Based Amplification (NASBA), Cleavase Fragment Length Polymorphism (U.S. Pat. No. 5,719,028), Isothermal and Chimeric Primer-initiated Amplification of Nucleic Acid (ICAN), Ramification-extension Amplification Method (RAM; U.S. Pat. Nos. 5,719,028 and 5,942,391) or other suitable methods for amplification of nucleic acids.

In order to amplify a nucleic acid with a small number of mismatches to one or more of the amplification primers, an amplification reaction may be performed under conditions of reduced stringency (e.g. a PCR amplification using an annealing temperature of 38°C, or the presence of 3.5 mM MgC12). The person skilled in the art will be able to select conditions of suitable stringency.

The primers herein are selected to be "substantially" complementary (i.e. at lea 65%, more preferably at least 80% perfectly complementary) to their target regions present on the different strands of each specific sequence to be amplified. It is possible to use primer sequences containing e.g. inositol residues or ambiguous bases or even primers that contain one or more mismatches when compared to the target sequence. I general, sequences that exhibit at least 65%, more preferably at least 80% homology with the target DNA or RNA oligonucleotide sequences, are considered suitable for use in a method of the present invention. Sequence mismatches are also not critical when using low stringency hybridization conditions.

The detection of the amplification products can in principle be accomplished by any suitable method known in the art. The detection fragments may be directly stainec or labelled with radioactive labels, antibodies, luminescent dyes, fluorescent dyes, or enzyme reagents. Direct DNA stains include for example intercalating dyes such as acridine orange, ethidium bromide, ethidium monoazide or Hoechst dyes. Alternatively, the DNA or RNA fragments may be detected by incorporation of labelled dNTP bases into the synthesized fragments. Detection labels which may be associated with nucleotide bases include e.g. fluorescein, cyanine dye or BrdUrd.

When using a probe-based detection system, a suitable detection procedure for use in the present invention may for example comprise an enzyme immunoassay (EIA) format (Jacobs et al., 1997, J. Clin. Microbiol. 35, 791-795). For performing a detection by manner of the EIA procedure, either the forward or the reverse primer used in the amplification reaction may comprise a capturing group, such as a biotin group for immobilization of target DNA PCR amplicons on e.g. a streptavidin coated microtiter plate wells for subsequent EIA detection of target DNA -amplicons (see below). The skilled person will understand that other groups for immobilization of target DNA PCR amplicons in an EIA format may be employed.

Probes useful for the detection of the target DNA as disclosed herein preferably bind only to at least a part of the DNA sequence region as amplified by the DNA amplification procedure. Those of skill in the art can prepare suitable probes for detection based on the nucleotide sequence of the target DNA without undue experimentation as set out herein. Also the complementary nucleotide sequences, whether DNA or RNA or chemically synthesized analogs, of the target DNA may suitably be used as type-specific detection probes in a method of the invention, provided that such a complementary strand is amplified in the amplification reaction employed.

Suitable detection procedures for use herein may for example comprise immobilization of the amplicons and probing the DNA sequences thereof by e.g. southern blotting. Other formats may comprise an EIA format as described above. To facilitate the detection of binding, the specific amplicon detection probes may comprise a label moiety such as a fluorophore, a chromophore, an enzyme or a radio-label, so as to facilitate monitoring of binding of the probes to the reaction product of the amplification reaction. Such labels are well-known to those skilled in the art and include, for example, fluorescein isothiocyanate (FITC), β-galactosidase, horseradish peroxidase, streptavidin, biotin, digoxigenin, 35S or 125I. Other examples will be apparent to those skilled in the art.

Detection may also be performed by a so called reverse line blot (RLB) assay, such as for instance described by Van den Brule et al. (2002, J. Clin. Microbiol. 40, 779-787). For this purpose RLB probes are preferably synthesized with a 5' amino group for subsequent immobilization on e.g. carboxyl-coated nylon membranes. The advantage of an RLB format is the ease of the system and its speed, thus allowing for high throughput sample processing.

The use of nucleic acid probes for the detection of RNA or DNA fragments is well known in the art. Mostly these procedure comprise the hybridization of the target nucleic acid with the probe followed by post-hybridization washings. Specificity is typically the function of post-hybridization washes, the critical factors being the ionic strength and temperature of the final wash solution. For nucleic acid hybrids, the Tm can be approximated from the equation of Meinkoth and Wahl, Anal. Biochem., 138: 267-284 (1984): Tm = 81.5 °C + 16.6 (log M) + 0.41 (% GC)-0.61 (% form)-500/L; where M is the molarity of monovalent cations, % GC is the percentage of guanosine and cytosine nucleotides in the nucleic acid, % form is the percentage of formamide in the hybridization solution, and L is the length of the hybrid in base pairs. The Tm is the temperature (under defined ionic strength and pH) at which 50% of a complementary target sequence hybridizes to a perfectly matched probe. Tm is reduced by about 1 °C for each 1 % of mismatching; thus, the hybridization and/or wash conditions can be adjusted to hybridize to sequences of the desired identity. For example, if sequences with > 90% identity are sought, the Tm can be decreased 10°C. Generally, stringent conditions are selected to be about 5 °C lower than the thermal melting point (Tm) for the specific sequence and its complement at a defined ionic strength and pH. However, severely stringent conditions can utilize a hybridization and/or wash at 1,2,3, or 4 °C

lower than the thermal melting point (Tm); moderately stringent conditions can utilize hybridization and/or wash at 6, 7, 8, 9, or 10 °C lower than the thermal melting point (Tm); low stringency conditions can utilize a hybridization and/or wash at 11, 12, 13, 14 15, or 20 °C lower than the thermal melting point (Tm). Using the equation, hybridization and wash compositions, and desired Tm, those of ordinary skill will understand that variations in the stringency of hybridization and/or wash solutions are inherently described. If the desired degree of mismatching results in a Tm of less than 45 °C (aqueous solution) or 32 °C (formamide solution) it is preferred to increase the SSC concentration so that a higher temperature can be used. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Laboratory Techniques in Biochemistry and Molecular Biology—Hybridization with Nucleic Acid Probes, Part I, Chapter 2" Overview of principles of hybridization and the strategy of nucleic acid probe assays", Elsevier. New York (1993); and Current Protocols in Molecular Biology, Chapter 2, Ausubel, et al., Eds., Greene Publishing and Wiley-Interscience, New York (1995).

In another aspect, the invention provides oligonucleotide probes for the generic detection of target RNA or DNA. The detection probes herein are selected to be "substantially" complementary to one of the strands of the double stranded nucleic acid generated by an amplification reaction of the invention. Preferably the probes are substantially complementary to the immobilizable, e.g. biotin labelled, antisense strand of the amplicons generated from the target RNA or DNA.

It is allowable for detection probes of the present invention to contain one or more mismatches to their target sequence. In general, sequences that exhibit at least 65%, more preferably at least 80% homology with the target oligonucleotide sequences are considered suitable for use in a method of the present invention.

Antibodies, both monoclonal and polyclonal, can also be used for detection purpose in the present invention, for example, in immunoassays in which they can be utilized in liquid phase or bound to a solid phase carrier. In addition, the monoclonal antibodies in these immunoassays can be detectably labeled in various ways. A variety of immunoassay formats may be used to select antibodies specifically reactive with a particular protein (or other analyte). For example, solid-phase ELISA immunoassays are routinely used to select monoclonal antibodies specifically immunoreactive with a protein. See Harlow and Lane, Antibodies, A Laboratory Manual, Cold Spring Harbor Publications, New York (1988), for a description of immunoassay formats and conditions that can be used to determine selective binding. Examples of types of immunoassays

that can utilize antibodies of the invention are competitive and non-competitive immunoassays in either a direct or indirect format. Examples of such immunoassays are the radioimmunoassay (RIA) and the sandwich (immunometric) assay. Detection of the antigens using the antibodies of the invention can be done utilizing immunoassays that are run in either the forward, reverse, or simultaneous modes, including immunohistochemical assays on physiological samples. Those of skill in the art will know, or can readily discern, other immunoassay formats without undue experimentation.

Antibodies can be bound to many different carriers and used to detect the presence of the target molecules. Examples of well-known carriers include glass, polystyrene, polypropylene, polyethylene, dextran, nylon, amylases, natural and modified celluloses, polyacrylamides, agaroses and magnetite. The nature of the carrier can be either soluble or insoluble for purposes of the invention. Those skilled in the art will know of other suitable carriers for binding monoclonal antibodies, or will be able to ascertain such using routine experimentation.

The invention also provides a method for serologically diagnosing an EMCR-CoV virus infection of a mammal comprising determining in a sample of said mammal the presence of an antibody specifically directed against an EMCR-CoV virus or component thereof by reacting said sample with a proteinaceous molecule or fragment thereof or an antigen according to the invention

Methods and means provided herein are particularly useful in a diagnostic kit for diagnosing an EMCR-CoV virus infection, be it by virological or serological diagnosis. Such kits or assays may for example comprise a virus, a nucleic acid, a proteinaceous molecule or fragment thereof, an antigen and/or an antibody according to the invention.

Use of a virus, a nucleic acid, a proteinaceous molecule or fragment thereof, an antigen and/or an antibody according to the invention is also provided for the production of a pharmaceutical composition, for example for the treatment or prevention of EMCR-CoV virus infections and/or for the treatment or prevention of atypical pneumonia, in particular in humans. Preferably a peptide comprising part of the amino acid sequence of the spike protein as depicted in the relevant translations of sequences E and F of figure 3, is used for the preparation of a therapeutic or prophylactic peptide. Also preferably, a protein comprising the amino acid sequence of the spike protein as depicted in the relevant translations of sequences E and F of figure 3, is used for the preparation of a sub-unit vaccine. Furthermore, the nucleocapsid of Coronaviruses, as

depicted in the translation of sequence F, in figure 3, is known to be particularly useful for eliciting cell-mediated immunity against Coronaviruses and can be used for the preparation of a sub-unit vaccine.

Attenuation of the virus can be achieved by established methods developed for this purpose, including but not limited to the use of related viruses of other species, serial passages through laboratory animals or/and tissue/cell cultures, serial passages through cell cultures at temparutes below 37°C (cold-adaption), site directed mutagenesis of molecular clones and exchange of genes or gene fragments between related viruses.

A pharmaceutical composition comprising a virus, a nucleic acid, a proteinaceous molecule or fragment thereof, an antigen and/or an antibody according to the invention can for example be used in a method for the treatment or prevention of an EMCR-CoV virus infection and/or a respiratory illness comprising providing an individual with a pharmaceutical composition according to the invention. This is most useful when said individual comprises a human. Antibodies against EMCR-CoV virus proteins, especially against the spike protein of EMCR-CoV virus, preferably against the amino acid sequence as depicted in translation 2 of sequence E and translation 1 of sequence F in figure 3, are also useful for prophylactic or therapeutic purposes, as passive vaccines. It is known from other coronaviruses that the spike protein is a very strong antigen and that antibodies against spike protein can be used in prophylactic and therapeutic vaccination.

The invention also provides method to obtain an antiviral agent useful in the treatment of atypical pneumonia comprising establishing a cell culture or experimental animal comprising a virus according to the invention, treating said culture or animal with an candidate antiviral agent, and determining the effect of said agent on said virus or its infection of said culture or animal. An example of such an antiviral agent comprises an EMCR-CoV virus-neutralising antibody, or functional component thereof, as provided herein, but antiviral agents of other nature are obtained as well.

The invention also provides use of an antiviral agent according to the invention for the preparation of a pharmaceutical composition, in particular for the preparation of a pharmaceutical composition for the treatment of atypical pneumonia, especifically when caused by an EMCR-CoV virus infection, and provides a pharmaceutical composition comprising an antiviral agent according to the invention, useful in a method for the treatment or prevention of an EMCR-CoV virus infection or atypical pneumonia,

said method comprising providing an individual with such a pharmaceutical composition.

The invention also comprises an animal model usable for testing of prophylactic and/or therapeutic methods and/or preparations. It is hypothesized that apes can be infected with the EMCR-CoV virus, thereby showing clinical symptoms, and more importantly, similar tissue morphology as found in humans suffering from atypical pneumonia caused by the EMCR-CoV virus. Subjecting apes to a prophylactic or therapeutic treatment either before or during infection with the virus will have a good and useful predictionary value for application of such a prophylaxis or therapy in human subjects.

The invention is further explained in the Examples without limiting it thereto.

Figure legends

Fig. 1: Phylogenetic relationship for the nucleotide sequences of isolate EMCR-CoV with its closest relatives genetically. Phylogenetic trees were generated by maximum likelihood analyses using 100 bootstraps and 3 jumbles. The scale representing the number of nucleotide changes is shown for each tree. Figure 1a. Maximum likelihood tree of matrix gene nucleotide sequences. Numbers in trees represent bootstrap values. The scale bar roughly reflects 10 % nucleotide differences between related sequences. Figure 1b. Maximum likelihood tree of nucleocapsid gene nucleotide sequences. Numbers in trees represent bootstrap values. The scale bar roughly reflects 10 % nucleotide differences between related sequences.

Fig. 2: Similarity matrix indicating the nucleotide and amino acid identity for the putative Matrix protein (2a and 2b resp.) and for the putavive Nucleoprotein (2c and 2d resp.) between the EMCR-CoV virus and closely related coronaviruses. See text for abbreviations.

Fig. 3: Nucleotide sequences from parts of the EMCR-CoV virus. Also included are the putative polypeptide sequences of polypeptides and alignments of the putative polypeptides with that of another member of the Coronoviridae family, where possible (mostly HCoV-229E).

Examples

Specimen collection

Virus was collected from an 8 month old patient suffering from pneumonia using nasal swabs.

Virus isolation and culture

Throat swabs were dipped into a culture of tMK cells and passaged four times. Virus was then in Vero-118 cells. One litre of virus containing cell culture supernatant was harvested, and the virus was pelleted in an ultracentrifuge and the virus pellet was resuspended in1ml PBS.

RNA isolation

RNA was isolated from the supernatant of infected cell cultures or sucrose gradient fractions using a High Pure RNA Isolation kit according to instructions from the manufacturer (Roche Diagnostics, Almere, The Netherlands).

Sequencing

Purified RNA was sent to BaseClear holding BV (Leiden, The Netherlands) for sequencing.

Phylogenetic analyses

Nucleotide sequences were aligned using Clustal W running under BioEdit version 5.0.9. Maximum likelihood trees were created using the Seqboot and DNA-ML packages of Phylip 5.6 using 100 bootstraps and 3 jumbles. The consensus trees were calculated using the Consense package of phylip 5.6. These consensus trees were used as usertree in DNA-ML to recalculate the branch lengths from the original sequences.

The sequences of EMCR-CoV were compared with those of reference viruses representing each species in the four groups of coronaviruses. These were: human coronavirus 229E (229E), af304460; porcine epidemic diarrhea virus (PEDV) af353511; transmissible gastroenteritis virus (TGEV), aj271965; bovine coronavirus (BoCoV), af220295; murine hepatitis virus (MHV), af201929; avian infectious bronchitis virus (AIBV), m95169, Canine coronavirus (CaCoV), d13096; feline coronavirus (FeCoV),

ay204704; porcine respiratory coronavirus (PRCoV), z24675; human coronavirus OC43 (OC43), m76373, l14643, m933990; porcine haemagglutinating encephalomyelitis virus (HEV), ay078417; rat coronavirus (RtCoV) af 207551) References for the viruses are the numbers of the NCBI catalog (http://www.ncbi.nlm.nih.gov/entrez/).

In general, coronaviruses, such as EMCR-CoV can be isolated and identified according to the following protocol:

Specimen collection

In order to find virus isolates nasopharyngeal aspirates, throat and nasal swabs, broncheo alveolar lavages, serum and plasma samples, and stools preferably from mammals such as humans, carnivores (dogs, cats, mustellits, seals etc.), horses, ruminants (cattle, sheep, goats etc.), pigs, rabbits, birds (poultry, ostriches, etc) should be examined. From birds cloaca swabs and droppings can be examined as well. Sera should be collected for immunological assays, such as ELISA, molecular-based assays, such as RT-PCR and virus neutralisation assays.

Collected virus specimens may be diluted with 5 ml Dulbecco MEM medium (BioWhittaker, Walkersville, MD) and thoroughly mixed on a vortex mixer for one minute. The suspension is thus centrifuged for ten minutes at 840 x g. The sediment is spread on a multispot slide (Nutacon, Leimuiden, The Netherlands) for immunofluorescence techniques, and the supernatant is used for virus isolation.

Virus isolation

For virus isolation Vero-118 cells or tMK cells (RIVM, Bilthoven, The Netherlands) were cultured in 24 well plates containing glass slides (Costar, Cambridge, UK), with the medium described below supplemented with 10% fetal bovine serum (BioWhittaker, Vervier, Belgium). Before inoculation the plates were washed with PBS and supplied with Eagle's MEM with Hanks' salt (ICN, Costa mesa, CA) supplemented with 0.52/liter gram $NaHCO_3$, 0.025 M Hepes (Biowhittaker), 2 mM L-glutamine (Biowhittaker), 200 units/liter penicilline, 200 µg/liter streptomycine (Biowhittaker), 1gram/liter lactalbumine (Sigma-Aldrich, Zwijndrecht, The Netherlands), 2.0 gram/liter D-glucose (Merck, Amsterdam, The Netherlands), 10 gram/liter peptone (Oxoid, Haarlem, The Netherlands) and 0.02% trypsine (Life Technologies, Bethesda, MD). The plates were inoculated with supernatant of the patient samples, 0,2 ml per well in triplicate, followed by centrifuging at 840x *g* for one hour. After inoculation the plates were

incubated at 37 °C for 1-7 days and cultures were checked daily for CPE. Extensive CPE was generally observed within 5-10 and included detachment of cells from the monolayer..

Virus culture

Sub-confluent monolayers of tMK cells or Vero clone 118 cells in media as described above were inoculated with supernatants of samples that displayed CPE or with samples taken from a patient.

RNA isolation

RNA was isolated from the supernatant of infected cell cultures or sucrose gradient fractions using a High Pure RNA Isolation kit according to instructions from the manufacturer (Roche Diagnostics, Almere, The Netherlands). RNA can also be isolated following other procedures known in the field (*Current Protocols in Molecular Biology*).

Sequence analysis

Sequence analyses were performed as follows: Purified viral RNA (500ng) was converted to cDNA using the SuperScript Choice system (Invitrogen Corp., Carlsbad, CA) by random priming according to the manufacturer's instructions. Blunt-ended, doublestranded cDNA fragments were size-selected on agarose gel to include fragments ranging from 750bp to 4kb. Following purification by spin column (Zymo Research, Orange, CA), cDNA fragments were ligated into pSMART-HCAmp (Lucigen Corp., Middleton, WI). The resulting library was electroporated into DH10B ElectroMAX cells (Invitrogen Corp., Carlsbad, CA), and inserts were amplified from individual colonies using pSMART AmpL1 and AmpR1 primers. PCR fragments were sequenced using BigDye 3.1 chemistry and run on a ABI3730 machine (Applied Biosystems, Foster City, CA).

Claims

1. An isolated essentially mammalian positive-sense single stranded RNA virus (EMCR-CoV) comprising one or more of the sequences of figure 3.

2. An isolated positive-sense single stranded RNA virus (EMCR-CoV) belonging to the Coronaviruses and identifiable as phylogenetically corresponding thereto by determining a nucleic acid sequence of said virus and testing it in phylogenetic tree analyses wherein maximum likelihood trees are generated using 100 bootstraps and 3 jumbles and finding it to be more closely phylogenetically corresponding to a virus isolate having the sequences as depicted in figure 3 than it is corresponding to a virus isolate of PEDV (porcine epidemic diarrhea virus), HCoV-229E (human coronavirus 229E), PRCoV (porcine respiratory coronavirus), TGEV (transmissible gastroenteritis virus), CaCoV (Canine coronavirus) and FeCoV (feline coronavirus).

3. A virus according to claim 1 or 2 wherein said nucleic acid sequence comprises a open reading frame (ORF) encoding a viral protein of said virus.

4. A virus according to claim 3 wherein said open reading frame is selected from th group of ORFs encoding the viral replicase, nucleocapsid protein, matrix protein or the spike protein.

5. A virus according to claim 1-4 isolatable from a human with respiratory tract disease such as, but not limited to, atypical pneumonia.

6. An isolated or recombinant nucleic acid or EMCR-CoV virus-specific functional fragment thereof obtainable from a virus according to anyone of claims 1 to 5.

7. A vector comprising a nucleic acid according to claim 6.

8. A host cell comprising a nucleic acid according to claim 6 or a vector according to claim 7.

9. An isolated or recombinant proteinaceous molecule or EMCR-CoV virus-specific functional fragment thereof encoded by a nucleic acid according to claim 6.

10. An antigen comprising a proteinaceous molecule or EMCR-CoV virus-specific functional fragment thereof according to claim 9.

11. An antibody specifically directed against an antigen according to claim 10.

12. A method for identifying a viral isolate as an EMCR-CoV virus comprising reacting said viral isolate or a component thereof with an antibody according to claim 11.

13. A method for identifying a viral isolate as an EMCR-CoV virus comprising reacting said viral isolate or a component thereof with a nucleic acid according to claim 6.

14. A method for virologically diagnosing an EMCR-CoV infection of a mammal comprising determining in a sample of said mammal the presence of a viral isolate or component thereof by reacting said sample with a nucleic acid according to claim 6 or an antibody according to claim 11.

15. A method for serologically diagnosing an EMCR-CoV infection of a mammal comprising determining in a sample of said mammal the presence of an antibody specifically directed against an EMCR-CoV virus or component thereof by reacting said sample with a proteinaceous molecule or fragment thereof according to claim 9 or an antigen according to claim 10.

16. A diagnostic kit for diagnosing an EMCR-CoV infection comprising a virus according to anyone of claims 1 to 5, a nucleic acid according to claim 6, a proteinaceous molecule or fragment thereof according to claim 9, an antigen according to claim 10 and/or an antibody according to claim 11.

17. Use of a virus according to any one claims 1 to 5, a nucleic acid according to claim 6, a vector according to claim 7, a host cell according to claim 8, a proteinaceous

molecule or fragment thereof according to claim 9, an antigen according to claim 10, or an antibody according to claim 11 for the production of a pharmaceutical composition.

18. Use according to claim 17 for the production of a pharmaceutical composition for the treatment or prevention of an EMCR-CoV virus infection.

19. Use according to claim 17 or 18 for the production of a pharmaceutical composition for the treatment or prevention of atypical pneumonia.

20. A pharmaceutical composition comprising a virus according to any one of claims 1 to 5, a nucleic acid according to claim 6, a vector according to claim 7, a host cell according to claim 8, a proteinaceous molecule or fragment thereof according to claim 9, an antigen according to claim 10, or an antibody according to claim 11.

21. A method for the treatment or prevention of an EMCR-CoV virus infection comprising providing an individual with a pharmaceutical composition according to claim 20.

22. A method for the treatment or prevention of atypical pneumonia comprising providing an individual with a pharmaceutical composition according to claim 20.

23. A viral replicase encoded by an RNA sequence comprising the sequences A, B, C, D, E and/or F, or homologues thereof as depicted in figure 3.

24. A viral spike protein comprising the amino acid sequence depicted as a translation of (part of) sequences E and F as depicted in figure 3, or a homologue thereof.

25 A viral nucleocapsid encoded by an RNA sequence comprising a translation of (part of) the sequence F as depicted in figure 3 or a homologue thereof.

26. A viral nsp 3 or envelope protein encoded by an RNA sequence comprising a translation of (part of) the sequence F as depicted in figure 3.

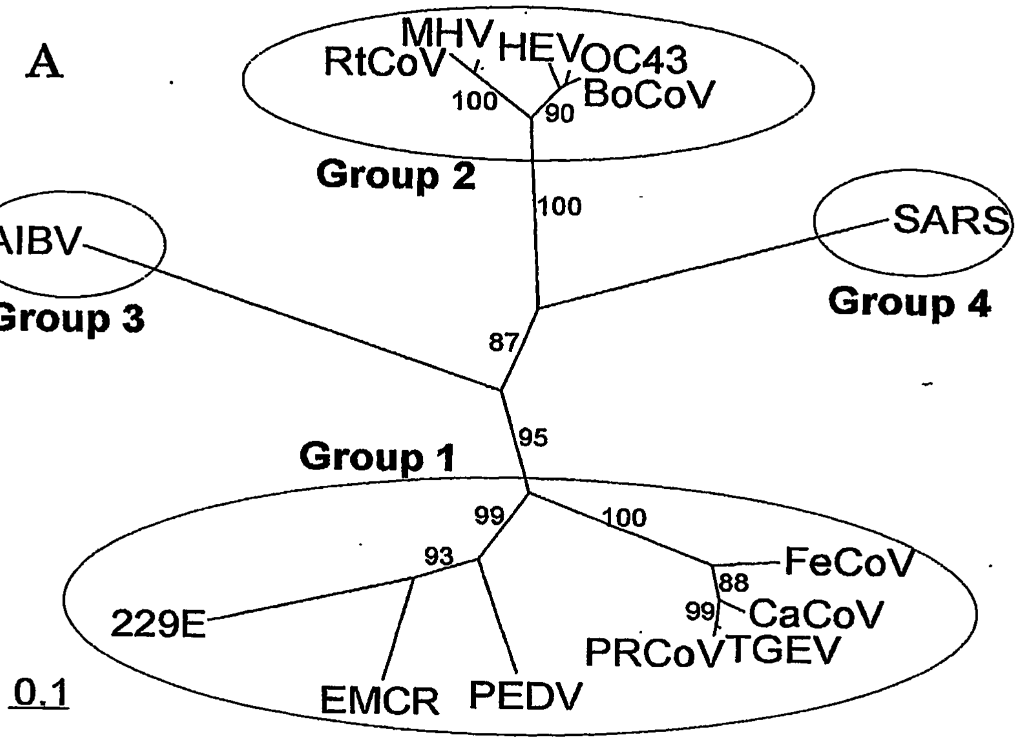
27. A nucleic acid sequence which comprises one or more of the sequences A to F as depicted in figure 3 or a nucleic acid sequence which can hybridise with any of these sequences under stringent conditions.

Abstract

The invention relates to the field of virology. The invention provides a new isolated essentially mammalian positive-sense single stranded RNA virus (EMCR-CoV) within the group of coronaviuses and components thereof.

Figure 1.

A


MHV
RtCoV
HEV
OC43
BoCoV
100
90
Group 2
100
AIBV
Group 3
SARS
Group 4
87
95
Group 1
99
100
93
FeCoV
88
99
CaCoV
229E
PRCoV
TGEV
0.1
EMCR
PEDV

B

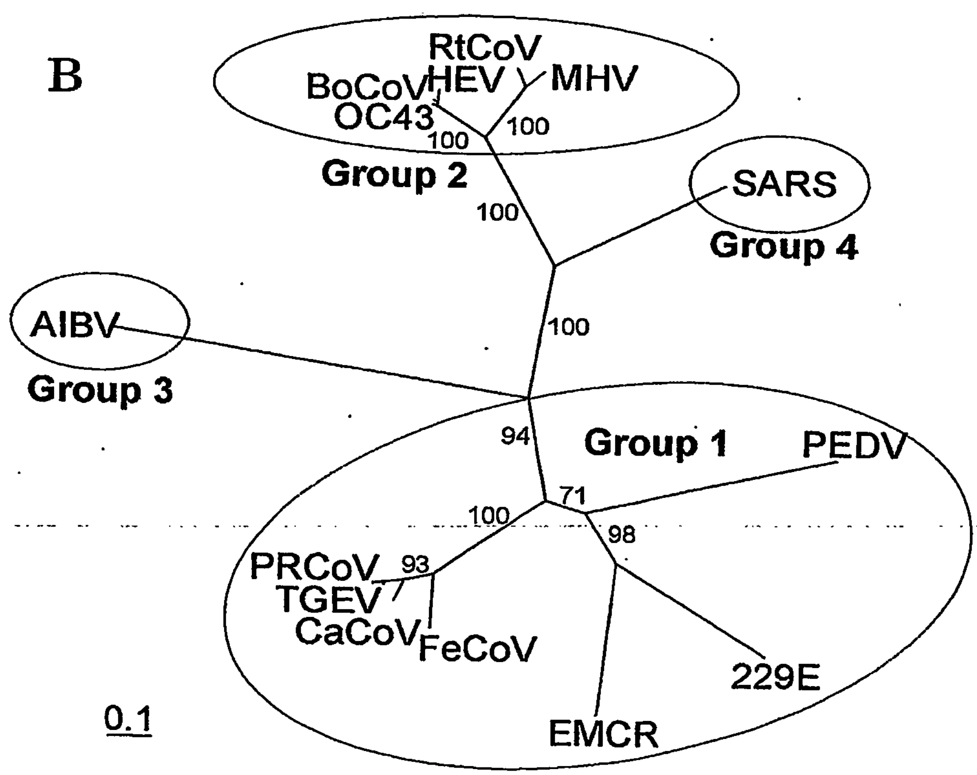

RtCoV
BoCoV
HEV
MHV
OC43
100
100
Group 2
SARS
100
Group 4
AIBV
100
Group 3
94
Group 1
PEDV
71
100
98
PRCoV
93
TGEV
CaCoV
FeCoV
229E
0.1
EMCR

Figure 2a

Seq->	SARS	EMCR	229E	PHDV	TGEV	CaCoV	FeCoV	PRCoV	OC43	PHEV	BoCoV	MHV	RatSA	AIBV
SARS	1.000	0.425	0.407	0.438	0.358	0.350	0.355	0.359	0.454	0.462	0.455	0.458	0.448	0.415
EMCR	—	1.000	0.632	0.650	0.467	0.461	0.479	0.462	0.471	0.465	0.465	0.444	0.437	0.429
229E	—	—	1.000	0.582	0.443	0.430	0.441	0.444	0.443	0.448	0.437	0.420	0.417	0.399
PEDV	—	—	—	1.000	0.485	0.476	0.469	0.482	0.448	0.460	0.453	0.451	0.451	0.408
TGEV	—	—	—	—	1.000	0.897	0.803	0.972	0.417	0.404	0.412	0.389	0.388	0.362
CaCoV	—	—	—	—	—	1.000	0.811	0.891	0.422	0.406	0.417	0.392	0.389	0.358
FeCoV	—	—	—	—	—	—	1.000	0.799	0.417	0.407	0.414	0.382	0.379	0.370
PRCoV	—	—	—	—	—	—	—	1.000	0.416	0.404	0.413	0.395	0.395	0.360
OC43	—	—	—	—	—	—	—	—	1.000	0.923	0.950	0.781	0.765	0.412
PHEV	—	—	—	—	—	—	—	—	—	1.000	0.920	0.775	0.757	0.407
BoCoV	—	—	—	—	—	—	—	—	—	—	1.000	0.791	0.777	0.398
MHV	—	—	—	—	—	—	—	—	—	—	—	1.000	0.930	0.401
RatSA	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.392
AIBV	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

Figure 2b

Seq->	SARS	EMCR	229E	PEDV	TGEV	CaCoV	FeCoV	PRCoV	OC43	PHEV	BoCoV	MHV	RatSA	AIBV
SARS	1.000	0.286	0.281	0.303	0.254	0.243	0.258	0.262	0.386	0.400	0.391	0.382	0.369	0.262
EMCR	—	1.000	0.615	0.650	0.441	0.429	0.441	0.437	0.317	0.317	0.317	0.303	0.303	0.239
229E	—	—	1.000	0.557	0.380	0.365	0.376	0.380	0.321	0.313	0.313	0.303	0.320	0.269
PEDV	—	—	—	1.000	0.460	0.452	0.425	0.460	0.351	0.360	0.364	0.358	0.363	0.234
TGEV	—	—	—	—	1.000	0.878	0.836	0.958	0.330	0.334	0.346	0.335	0.332	0.208
CaCoV	—	—	—	—	—	1.000	0.835	0.878	0.311	0.315	0.326	0.319	0.304	0.192
FeCoV	—	—	—	—	—	—	1.000	0.851	0.296	0.307	0.315	0.300	0.292	0.192
PRCoV	—	—	—	—	—	—	—	1.000	0.330	0.334	0.346	0.335	0.332	0.215
OC43	—	—	—	—	—	—	—	—	1.000	0.934	0.947	0.848	0.818	0.270
PHEV	—	—	—	—	—	—	—	—	—	1.000	0.943	0.848	0.818	0.270
BoCoV	—	—	—	—	—	—	—	—	—	—	1.000	0.870	0.839	0.278
MHV	—	—	—	—	—	—	—	—	—	—	—	1.000	0.938	0.271
RatSA	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.275
AIBV	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

Figure 2c

Seq->	SARS	EMCR	229E	PEDV	TGEV	FeCoV	PRCoV	CaCoV	RSDAC	MHV	PHEV	OC43	BoCoV	AIBV
SARS	1,000	0,303	0,281	0,281	0,343	0,325	0,340	0,337	0,430	0,431	0,459	0,457	0,459	0,304
EMCR	—	1,000	0,518	0,388	0,450	0,442	0,453	0,450	0,291	0,296	0,300	0,306	0,302	0,331
229E	—	—	1,000	0,409	0,441	0,424	0,443	0,437	0,295	0,295	0,290	0,297	0,295	0,340
PEDV	—	—	—	1,000	0,374	0,356	0,373	0,377	0,271	0,275	0,269	0,269	0,267	0,318
TGEV	—	—	—	—	1,000	0,782	0,965	0,897	0,335	0,334	0,323	0,325	0,328	0,340
FeCoV	—	—	—	—	—	1,000	0,776	0,778	0,327	0,320	0,313	0,318	0,316	0,327
PRCoV	—	—	—	—	—	—	1,000	0,879	0,336	0,331	0,325	0,325	0,329	0,347
CaCoV	—	—	—	—	—	—	—	1,000	0,324	0,328	0,319	0,321	0,322	0,343
RSDAC	—	—	—	—	—	—	—	—	1,000	0,891	0,698	0,707	0,707	0,297
MHV	—	—	—	—	—	—	—	—	—	1,000	0,692	0,704	0,701	0,288
PHEV	—	—	—	—	—	—	—	—	—	—	1,000	0,949	0,954	0,296
OC43	—	—	—	—	—	—	—	—	—	—	—	1,000	0,971	0,298
BoCoV	—	—	—	—	—	—	—	—	—	—	—	—	1,000	0,296
AIBV	—	—	—	—	—	—	—	—	—	—	—	—	—	1,000

Figure 2d

Seq->	EMCR	229E	PEDV	TGEV	FeCoV	PRCoV	CaCoV	RSDAC	MHV	PHEV	OC43	BoCoV	SARS	AIBV
EMCR	1,000	0,447	0,358	0,336	0,326	0,334	0,344	0,188	0,189	0,179	0,183	0,183	0,210	0,173
229E	—	1,000	0,336	0,335	0,304	0,328	0,333	0,196	0,204	0,187	0,190	0,188	0,199	0,173
PEDV	—	—	1,000	0,277	0,244	0,272	0,270	0,163	0,168	0,160	0,160	0,158	0,184	0,178
TGEV	—	—	—	1,000	0,761	0,963	0,897	0,220	0,223	0,200	0,202	0,200	0,232	0,192
FeCoV	—	—	—	—	1,000	0,756	0,763	0,209	0,212	0,185	0,187	0,189	0,218	0,185
PRCoV	—	—	—	—	—	1,000	0,879	0,220	0,228	0,202	0,204	0,202	0,230	0,192
CaCoV	—	—	—	—	—	—	1,000	0,215	0,221	0,196	0,198	0,196	0,216	0,196
RSDAC	—	—	—	—	—	—	—	1,000	0,894	0,693	0,697	0,697	0,285	0,200
MHV	—	—	—	—	—	—	—	—	1,000	0,680	0,684	0,682	0,282	0,208
PHEV	—	—	—	—	—	—	—	—	—	1,000	0,948	0,953	0,261	0,195
OC43	—	—	—	—	—	—	—	—	—	—	1,000	0,973	0,261	0,197
BoCoV	—	—	—	—	—	—	—	—	—	—	—	1,000	0,266	0,197
SARS	—	—	—	—	—	—	—	—	—	—	—	—	1,000	0,211
AIBV	—	—	—	—	—	—	—	—	—	—	—	—	—	1,000

Figure 3

RNA sequences, implied polypeptides and alignment with one close relative

1. Sequence A

3762 Nucleotides encoding part of Replicase

ATTCGTTCTATAGATAGAGAATTTTCTTATTTAGACTTTGTGTCTACTCCTCTCAACTAAACGAAATTTTTCTAG
TGCTGTCATTTGTTATGGCAGTCCTAGTGTAATTGAAATTTCGTCAAGTTTGTAAACTGGTTAGGCAAGTGTTGT
ATTTTCTGTGTTTAAGCACTGGTGGTTCTGTCCACTAGTGCACACATTGATACTTAAGTGGTGTTCTGTCACTGC
TTATTGTGGAAGCAACGTTCTGTCGTTGTGGAAACCAATAACTGCTAACCATGTTTTACAATCAAGTGACACTTG
CTGTTGCAAGTGATTCGGAAATTTCAGGTTTTGGTTTTGCCATTCCTTCTGTAGCCGTTCGCGCTTATAGCGAAG
CCGCTGCACAAGGTTTTCAGGCATGCCGCTTTGTTGCTTTTGGCTTACAGGATTGTGTAACCGGTATTAATGATG
ACGATTATGTCATTGCATTGACTGGTACTAATCAGCTTTGTGCCAAAATTTTACTTTTTTCTGATAGACCTCTTA
ATTTGCGAGGTTGGCTCATTTTTTCTAACAGCAATTATGTTCTTCAGGACTTTGATGTTGTTTTTGGCCATGGTG
CAGGAAGTGTGGTTTTTGTGGATAAGTATATGTGTGGTTTTGATGGTAAACCTGTGTTACCTAAAAACATGTGGG
AATTTAGAGATTACTTTAATGATAATACTGATAGTATTGTTATTGGTGGTGTCACTTATCAATTAGCATGGGATG
TTATACGTAAAGACCTTTCTTATGAACAGCAAAATGTTTTAGCTATTGAGAGCATTCATTATCTTGGCACTACAG
GTCATACTTTGAAGTCTGGTTGCAAACTCATTAATGCCAAGCCGCCTAAATATTCTTCTAAGGTTGTTTTGAGTG
GTGAATGGAATGCTGTGTATAAGGCGTTTGGTTCACCATTTATTACAAATGGTATATCATTGCTAGATATAATTG
TTAAACCAGTTTTCTTTAATGCTTTTGTTAAATGCAATTGTGGTTCTGAGAATTGGAGTGTTGGTGCATGGGATG
GTTATCTATCTTCTTGTTGTGGCACACCTGCTAAGAAACTTTGTGTTGTTCCTGGTAATGTTGTTCCTGGTGATG
TGATCATCACCTCAACTGATGCTGGTTGTGGTGTTAAATACTATGCTGGCTTAGTTGTTAAACATATTACTAACA
TTACTGGTGTGTCTTTATGGCGTGTTACAGCTGTTCATTCTGATGGAATGTTTGTGGCAACATCTTCTTATGATG
CACTTTTGCATAGAAATTCATTAGACCCTTTTTGCTTTGATGTTAACACTTTACTTTCTAATCAATTACGTCTAG
CTTTTCTTGGTGCTTCTGTTACAGAAGATGTTAAATTTGCTGCTAGCACTGGTGTTATTGACATTAGTGCTGGTA
TGTTTGGTCTTTACGATGACATATTGACAAACAATAAACCTTGGTTTGTACGCAAAGCTTCTGGGCTTTTTGATG
CAATCTGGGATGCTTTTGTTGCCGCTATTAAGCTTGTGCCAACTACTACTGGTGGTTTGGTTAGGTTTGTTAAGT
CTATCGCTTCAACTGTTTTAACTGTTTCTAATGGTGTTATTATTATGTGTGCAGATGTTCCAGATGCTTTTCAAC
CAGTTTACCGCACATTTACACAAGCTATTTGTGCTGCATTTGATTTTTCTTTAGATGTATTTAAAATTGGTGATG
TTAAATTTAAACGACTTGGTGATTATGTTCTTACTGAAAATGCTCTTGTTCGTTTGACTACTGAAGTTGTTCGTG
GTGTTCGTGATGCTCGCATAAAGAAAGCCATGTTTACTAAAGTAGTTGTAGGTCCTACAACTGAAGTTAAGTTTT
CTGTTATTGAACTTGCCACTGTTAATTTGCGTCTTGTTGATTGTGCACCTGTAGTTTGCCCTAAAGGTAAAATTG
TTGTTATTGCTGGACAAGCTTTTTTCTATAGTGGTGGTTTTTATCGTTTTATGGTTGATTCTACAACTGTATTAA
ATGACCCTGTTTTTACTGGTGAGTTATTTTATACTATTAAGTTTAGTGGTTTTAAGCTTGATGGTTTTAACCATC
AGTTTGTTAATGCTAGTTCTGCTACAGATGCCATTATTGCTGTTGAGCTGTTGTTATCGGATTTTAAAACTGCAG
TTTTTGTGTACACATGTGTGGTTGATGGTTGTAGTGTCATTGTTAGACGTGATGCTACATTCGCCACACATGTGT
GTTTTAAGGACTGTTATAGTATTTGGGAGCAATTCTGCATTGATAATTGTGGTGAGCCATGGTTTTTGACTGATT
ATAATGCTATCTTGCAGAGTAATAACCCTCAATGTGCTATTGTTCAAGCATCGGAGTCTAAAGTTTTGCTTGAGA
GGTTTTTACCTAAGTGTCCTGAAGTACTGTTGAGTATTGATGATGGCCATTTATGGAATCTTTTTGTTGAAAAGT
TTAATTTTGTTACAGATTGGTTAAAAACTCTTAAGCTTACACTTACTTCTAATGGTCTTTTAGGTAATTGTGCCA
AACGTTTTAGACGTGTTTTGGTAAAATTGCTTGATGTCTATAATGGTTTTCTTGAAACTGTCTGTAGTGTCGTAC
ACACTGCTGGTGTTTGCATTAAATATTATGCTGTTAATGTTCCATATGTAGTTATTAGTGGTTTTGTAAGTCGTG
TAATTCGTAGAGAAAGGTGTGACGTGACTTTTCCTTGTGTTAGTTGTGTCACTTTTTTCTATGAATTTTTAGACA
CGTGTTTTGGTGTTAGTAAACCTAATGCCATTGATGTTGAACATTTAGAGCTTAAAGAAACTGTTTTTGTTGAAC
CTAAGGATGGTGGTCAATTTTTTGTTTCTGATGATTATCTTTGGTATGTTGTAGATGACATTTATTATCCAGCTT
CATGTAATGGTGTATTGCCAGTTGCTTTTACAAAATTGGCAGGTGGTAAAATATCTTTTTCTGATGATGTTATAG
TTCATGATGTTGAACCTACCCATAAAGTCAAGCTCATATTTGAGTTTGAAGATGATGTTGTTACCAGTCTTTGTA
AGAAGAGTTTTGGTAAGTCTATTATTTATACAGGTGATTGGGAAGGTTTACATGAAGTTCTTACATCTGCAATGA
ATGTCATTGGGCAACATATTAAGTTGCCACAATTTTATATTTATGATGAAGAGGGTGGTTATGATGTTTCTAAAC
CAGTTATGATTTCACAATGGCCTATTAGTGATGATAGTGATGGTTGTGTTGTTGAAGCGAGCACTGATTTTCATC
AATTAGAATCTGTTAGAGAAGAGGTTGATATAATTGAACAACCTTTTGGGGAAGTTGAACATGCGCTCTCAATTA
GACAACCTTTTTCTTTTTCTTTTAGAGATGAATTGGGTGTTCGTGTTTTAGATCAATCTGATAATAATTGTTGGA
TTAGTACCACACTTATACAGTTGCAACTTACAAAGCTTTTGGATGATTCTATTGAGATGCAATTGTTTAAAGTTG
GTAAAGTTGATTCAATTGTTCAAAAGTGTTATGAGTTGTCTCATTTAATTAGTGGTTCACTTGGTGATAGTGGTA
AACTTCTTAGTGAACTTCTTAAAGATAAATATACATGTTCTATAACTTTTGAGATGTCTTGTGATTGTGGTAAAA
AGTTTGATGAGCAAGTTGGTTGTTTGTTTTGGATTATGCCTTACACAAAACTTTTCAAAAAGGTGAGAACGAATT
CAGCTGTTCTCG

Putative ORFs

>~out: 140 to 310: Frame 2 57 aa
ASVVFSVFKHWWFCPLVHTLILKWCSVTAYCGSNVLSLWKPITANHVLQSSDTCCCK
>~out: 267 to 3761: Frame 3 1165 aa

```
LLTMFYNQVTLAVASDSEISGFGFAIPSVAVRAYSEAAAQGFQACRFVAFGLQDCVTGINDDDYVIALTGTNQL
AKILLFSDRPLNLRGWLIFSNSNYVLQDFDVVFGHGAGSVVFVDKYMCGFDGKPVLPKNMWEFRDYFNDNTDSI
IGGVTYQLAWDVIRKDLSYEQQNVLAIESIHYLGTTGHTLKSGCKLINAKPPKYSSKVVLSGEWNAVYKAFGSP
ITNGISLLDIIVKPVFFNAFVKCNCGSENWSVGAWDGYLSSCCGTPAKKLCVVPGNVVPGDVIITSTDAGCGVK
YAGLVVKHITNITGVSLWRVTAVHSDGMFVATSSYDALLHRNSLDPFCFDVNTLLSNQLRLAFLGASVTEDVKF
ASTGVIDISAGMFGLYDDILTNNKPWFVRKASGLFDAIWDAFVAAIKLVPTTTGGLVRFVKSIASTVLTVSNGV
IMCADVPDAFQPVYRTFTQAICAAFDFSLDVFKIGDVKFKRLGDYVLTENALVRLTTEVVRGVRDARIKKAMFT
VVVGPTTEVKFSVIELATVNLRLVDCAPVVCPKGKIVVIAGQAFFYSGGFYRFMVDSTTVLNDPVFTGELFYTI
FSGFKLDGFNHQFVNASSATDAIIAVELLLSDFKTAVFVYTCVVDGCSVIVRRDATFATHVCFKDCYSIWEQFC
DNCGEPWFLTDYNAILQSNNPQCAIVQASESKVLLERFLPKCPEVLLSIDDGHLWNLFVEKFNFVTDWLKTLKL
LTSNGLLGNCAKRFRRVLVKLLDVYNGFLETVCSVVHTAGVCIKYYAVNVPYVVISGFVSRVIRRERCDVTFPC
SCVTFFYEFLDTCFGVSKPNAIDVEHLELKETVFVEPKDGGQFFVSDDYLWYVVDDIYYPASCNGVLPVAFTKL
GGKISFSDDVIVHDVEPTHKVKLIFEFEDDVVTSLCKKSFGKSIIYTGDWEGLHEVLTSAMNVIGQHIKLPQFY
YDEEGGYDVSKPVMISQWPISDDSDGCVVEASTDFHQLESVREEVDIIEQPFGEVEHALSIRQPFSFSFRDELG
RVLDQSDNNCWISTTLIQLQLTKLLDDSIEMQLFKVGKVDSIVQKCYELSHLISGSLGDSGKLLSELLKDKYTC
ITFEMSCDCGKKFDEQVGCLFWIMPYTKLFKKVRTNSAVL
>~out: 472 to 738: Frame 1        89 aa
LVLISFVPKFYFFLIDLLICEVGSFFLTAIMFFRTLMLFLAMVQEVWFLWISICVVLMVNLCYLKTCGNLEITL
IILIVLLLVVSLIN
>~out: 973 to 1125: Frame 1        51 aa
LLNQFSLMLLLNAIVVLRIGVLVHGMVIYLLVVAHLLRNFVLFLVMLFLVM
>~out: 2026 to 2316: Frame 1        97 aa
MTLFLLVSYFILLSLVVLSLMVLTISLLMLVLLQMPLLLLSCCYRILKLQFLCTHVWLMVVVSLLDVMLHSPHM
VLRTVIVFGSNSALIIVVSHGF
```

Alignment

>gi|281286|pir||S28600 hypothetical protein 1a - human coronavirus
gi|59491|emb|CAA49877.1| ORF1a [Human coronavirus 229E]
Length = 4085

Score = 882 bits (2280), Expect = 0.0
Identities = 470/1159 (40%), Positives = 675/1159 (58%), Gaps = 7/1159 (0%)
Frame = +3

```
Query: 276  MFYNQVTLAVASDSEISGFGFAIPSVAVRAYSEAAAQGFQACRFVAFGLQDCVTGINDDD 455
            M  N+VTLAVASDSEIS  G +  + AVR YSEAA+ GF+ACRFV+  LQDC+ GI DD
Sbjct: 1    MACNRVTLAVASDSEISANGCSTIAQAVRRYSEAASNGFRACRFVSLDLQDCIVGIADDT 60

Query: 456  YVIALTGTNQLCAKILLFSDRPLNLRGWLIFSNSNYVLQDFDVVFG-HGAGSVVFVDKYM 632
            YV+ L G   L   I+ FSDRP  L GWL+FSNSNY+L++FDVVFG  G G+V + D+Y+
Sbjct: 61   YVMGLHGNQTLFCNIMKFSDRPFMLHGWLVFSNSNYLLEEFDVVFGKRGGGNVTYTDQYL 120

Query: 633  CGFDGKPVLPKNMWEFRDYFNDNTDSIVIGGVTYQLAWDVIRKDLSYEQQNVLAIESIHY 812
            CG DGKPV+ +++W+F D+F +N + I+I G TY  AW   RK L Y++QN LAIE I Y
Sbjct: 121  CGADGKPVMSEDLWQFVDHFGEN-EEIIINGHTYVCAWLTKRKPLDYKRQNNLAIEEIEY 179

Query: 813  L-GTTGHTLKSGCKLINAKPPKYSSKVVLSGEWNAVYKAFGSPFITNGISLLDIIVKPVF 989
            + G   HTL++G  L  AK  K SSKVVLS    + +YK FGSP +TNG ++L+    KPVF
Sbjct: 180  VHGDALHTLRNGSVLEMAKEVKTSSKVVLSDALDKLYKVFGSPVMTNGSNILEAFTKPVF 239

Query: 990  FNAFVKCNCGSENWSVGAWDGYLSSCCGTPAKKLCVVPGNVVPGDVIITSTDAGCGVKYY 1169
             +A V+C CG+++WSVG W G+ SSCC    + KLCVVPGNV PGD +IT+  AG G+KY+
Sbjct: 240  ISALVQCTCGTKSWSVGDWTGFKSSCCNVISNKLCVVPGNVKPGDAVITTQQAGAGIKYF 299

Query: 1170 AGLVVKHITNITGVSLWRVTAVHSDGMFVATSSYDALLHRNSLDPFCFDVNTLLSNQLRL 1349
             G+ +K + NI GVS+WRV A+ S   FVA+S++    H N +D FCF+V    ++++RL
Sbjct: 300  CGMTLKFVANIEGVSVWRVIALQSVDCFVASSTFVEEEHVNRMDTFCFNVRNSVTDECRL 359

Query: 1350 AFLGASVTEDVKFAASTGVIDISAGMFGLYDDILTNNKPWFVRKASGLFDAIWDAFVAAI 1529
            A LGA +T +V+    ++GVIDIS G F +YDDI   +KPWFVRKA +F   W A  +A+
Sbjct: 360  AMLGAEMTSNVRRQVASGVIDISTGWFDVYDDIFAESKPWFVRKAEDIFGPCWSALASAL 419

Query: 1530 KLVPTTTGGLVRFVKSIASTVLTVSNGVIIMCADVPDAFQPVYRTFTQAICAAFDFSLDV 1709
            K +   TTG LVRFVKSI ++ + V  G I + A VP+ F   +  F AI  FD +++
Sbjct: 420  KQLKVTTGELVRFVKSICNSAVAVVGGTIQILASVPEKFLNAFDVFVTAIQTVFDCAVET 479

Query: 1710 FKIGDVKFKRLGDYVLTENALVRLTTEVVRGVRDARIKKAMFTKVVVGPTTEVKFSVIEL 1889
              I    F ++ DYVL +NALV+L T  ++GVR+  + K  +  VVVG T EVK S +E
Sbjct: 480  CTIAGKAFDKVFDYVLLDNALVKLVTTKLKGVRERGLNKVKYATVVVGSTEEVKSSRVER 539
```

Fig 3. (cont)

```
Query: 1890 ATVNLRLVDCAPVVCPKGKIVVIAGQAFFYSGGFYRFMVDSTTVLNDPVFTGELFYTIKF 2069
            +T  L + +    +  +G  VVI   A+F S G++R M    +VL   V+     + +
Sbjct: 540  STAVLTIANNYSKLFDEGYTVVIGDVAYFVSDGYFRLMASPNSVLTTAVYKPLFAFNVNV 599

Query: 2070 SGFKLDGFNHQFVNASSATDAIIAVELLLSDFKTAVFVYTCVVDGCSVIVRRDATFATHV 2249
             G + + F    V   +   A++ V   +++F+     Y+  V    +IV+ + +    +
Sbjct: 600  MGTRPEKF-PTTVTCENLESAVLFVNDKITEFQ---LDYSIDVIDNEIIVKPNISLCVPL 655

Query: 2250 CFKDCYSIWEQFCIDNCGEPWFLTDYNAILQSNNPQCAIVQASESKVLLERFLPKCPEVL 2429
              +D    W+ FC     E WF  DY A +   +   A V+A+ESK  ++  +P CP +L
Sbjct: 656  YVRDYVDKWDDFCRQYSNESWFEDDYRAFISVLDITDAAVKAAESKAFVDTIVPPCPSIL 715

Query: 2430 LSIDDGHLWNLFVEKFNFVTDWXXXXXXXXXXXXXXXXCAKRFRRVLVKLLDVYNGFLET 2609
              ID G +WN  ++  N V DW                CAKRF+R L  LL+ YN FL+T
Sbjct: 716  KVIDGGKIWNGVIKNVNSVRDWLKSLKLNLTQQGLLGTCAKRFKRWLGILLEAYNAFLDT 775

Query: 2610 VCSVVHTAGVCIKYYAVNVPYVVISGFVSRVIRRERCD--VTFPCVSCVTFFYEFLDTCF 2783
            V S V   G+  K YA + PY+VI   V +V  +   +    FP    +  F  F
Sbjct: 776  VVSTVKIGGLTFKTYAFDKPYIVIRDIVCKVENKTEAEWIELFPHNDRIKSFSTFESAYM 835

Query: 2784 GVSKPNAIDVEHLELKETVFVEPKDGGQFFVSDDYLWYVVDDIYYPASCNGVLPVAFTKL 2963
             ++ P   D+E +EL +  FVEP  GG   V D++++Y  D +YYP++   +LPVAFTK
Sbjct: 836  PIADPTHFDIEEVELLDAEFVEPGCGGILAVIDEHVFYKKDGVYYPSNGTNILPVAFTKA 895

Query: 2964 AGGKISFSDDVIVHDVEPTHKVKLIFEFEDDVVTSLCKKSFGKSIIYTGDWEGLHEVLTS 3143
            AGGK+SFSDDV V D+EP ++VKL FEFED+ +  +C+K+ GK I + GDW+    + + S
Sbjct: 896  AGGKVSFSDDVEVKDIEPVYRVKLCFEFEDEKLVDVCEKAIGKKIKHEGDWDSFCKTIQS 955

Query: 3144 AMNVIGQHIKLPQFYIYDEEGGYDVSKPVMISQWPIS---DDSDGCVVEASTDFHQLESV 3314
            A++V+  ++ LP +YIYDEEGG D+S PVMIS+WP+S      +  + + + D   ++ V
Sbjct: 956  ALSVVSCYVNLPTYYIYDEEGGNDLSLPVMISEWPLSVQQAQQEATLPDIAEDV--VDQV 1013

Query: 3315 REEVDIIEQPFGEVEHALSIRQPFSFSFRDELGVRVLDQSDNNCWISXXXXXXXXXXXXD 3494
             E   I +    +V+H +S   PF   F +  G+++L Q DNNCW++            D
Sbjct: 1014 EEVNSIFDIETVDVKHDVS---PFEMPFEELNGLKILKQLDNNCWVNSVMLQIQLTGILD 1070

Query: 3495 DSIEMQLFKVGKVDSIVQKCYELSHLIXXXXXXXXXXXXXXXXXXYTCSITFEMSCDCGK 3674
                MQ FK+G+V  ++++CY     I                  +T  +  +  C C
Sbjct: 1071 GDYAMQFFKMGRVAKMIERCYTAEQCIRGAMGDVGLCMYRLLKDLHTGFMVMDYKCSCTS 1130

Query: 3675 KFDEQVGCLFWIMPYTKLF 3731
               E+ G + +  P  K F
Sbjct: 1131 GRLEESGAVLFCTPTKKAF 1149
```

2. Sequence B

1610 nucleotides encodes part of replicase

TTTCTGCCTATGGAGGTCAGGTATGATTTAAATGGTCAGTATTGAGCGATATCTAGAGAATTCGTCTGAAAATGG
TATTCCACTTATGCCTCTTCTTAGTTGTGGTATTTTTGGTGTAAGGATTGAAAATTCTCTTAAAGCTTTGTTTAG
TTGTGACATTAATAAACCATTGCAAGTTTTTGTTTATTCTTCAAATGAAGAACAAGCTGTTCTTAAGTTTTTAGA
TGGTTTAGATTTAACACCAGTCATTGACGATGTTGATGTTGTTAAACCTTTTAGAGTTGAAGGTAATTTTTCATT
CTTTGATTGTGGTGTCAATGCCTTGGATGGTGATATTTACTTATTATTTACTAACTCTATTTTAATGTTGGATAA
ACAAGGACAATTATTGGACACAAAACTTAATGGTATTTTGCAACAGGCAGTTCTTGATTATCTTGCTACAGTTAA
AACTGTACCAGCTGGTAATTTGGTTAAACTTGTTGTTGAGAGTTGTACCATTTATATGTGTGTTGTACCATCGAT
AAATGATCTTTCTTTTGATAAAAATCTTGGTCGTTGTGTGCGTAAACTTAATAGATTGAAAACTTGTGTTATTGC
CAATGTTCCTGCTATTGATGTTTTGAAAAAGCTTCTTTCAAGTTTGACTTTAACTGTTAAATTTGTTGTAGAGAG
TAATGTTATGGATGTTAACGACTGTTTTAAGAATGATAATGTAGTTTTGAAAATTACTGAAGATGGTATTAATGT
TAAAGATGTTGTTGTTGAGTCTTCTAAGTCACTTGGTAAACAATTGGGTGTTGTGAGTGATGGTGTTGACTCTTT
TGAAGGTGTTTTACCTATTAATACTGATACTGTCTTATCTGTAGCTCCAGAAGTTGACTGGGTTGCTTTTTACGG
TTTTGAAAAGGCAGCACTTTTTGCTTCTTTGGATGTAAAGCCATATGGTTACCCTAATGATTTTGTTGGTGGTTT
TAGAGTTCTTGGGACCACCGACAATAATTGTTGGGTTAATGCAACTTGTATAATTTTACAGTATCTTAAGCCTAC
TTTTAAATCTAAGGGTTTAAATGTTCTTTGGAACAAATTTGTTACAGGTGATGTTGGACCTTTTGTTAGTTTTAT
TTATTTTATAACTATGTCTTCAAAGGGTCAAAAGGGTGATGCTGAAGAGGCATTATCTAAATTGTCAGAGTATTT
GATTAGTGATTCTATTGTTACTCTTGAACAATATTCAACTTGTGACATTTGTAAAAGTACTGTAGTTGAAGTTAA
AAGTGCTGTTGTCTGTGCTAGTGTGCTTAAAGATGGTTGTGATGTTGGTTTTTGTCCACACAGACATAAATTGCG
TTCACGTGTTAAGTTTGTTAATGGACGTGTTGTTATTACCAATGTTGGTGAACCTATAATTTCACAACCTTCTAA
GTTGCTTAATGGTATTGCTTATACAACATTTTCAGGTTCTTTTGATAACGGTCACTATGTAGTTTATGATGCTGC
TAATAATGCTGTCTATGATGGTGCTCGTTTATTTGCTTCAGATTTGTCTACTTTAGCTGTTACAGCTATTGTTGT
AGTAGGTGGTTGTGTAACATCTAATTTCCACAACG

Fig 3. (cont) 9/25

Putative ORFs

>~out: 32 to 1609: Frame 2 526 aa
MVSIERYLENSSENGIPLMPLLSCGIFGVRIENSLKALFSCDINKPLQVFVYSSNEEQAVLKFLDGLDLTPVID
VDVVKPFRVEGNFSFFDCGVNALDGDIYLLFTNSILMLDKQGQLLDTKLNGILQQAVLDYLATVKTVPAGNLVK
VVESCTIYMCVVPSINDLSFDKNLGRCVRKLNRLKTCVIANVPAIDVLKKLLSSLTLTVKFVVESNVMDVNDCF
NDNVVLKITEDGINVKDVVVESSKSLGKQLGVVSDGVDSFEGVLPINTDTVLSVAPEVDWVAFYGFEKAALFAS
DVKPYGYPNDFVGGFRVLGTTDNNCWVNATCIILQYLKPTFKSKGLNVLWNKFVTGDVGPFVSFIYFITMSSKG
KGDAEEALSKLSEYLISDSIVTLEQYSTCDICKSTVVEVKSAVVCASVLKDGCDVGFCPHRHKLRSRVKFVNGR
VITNVGEPIISQPSKLLNGIAYTTFSGSFDNGHYVVYDAANNAVYDGARLFASDLSTLAVTAIVVVGGCVTSNF
N
>~out: 366 to 524: Frame 3 53 aa
CWINKDNYWTQNLMVFCNRQFLIILLQLKLYQLVIWLNLLLRVVPFICVLYHR

Alignment

>gi|12175747|ref|NP_073549.1| replicase polyprotein 1ab [Human coronavirus 229E]
gi|30179827|sp|Q05002|R1AB_CVH22 Replicase polyprotein 1ab (pp1ab) (ORF1ab polyprotein) [Includes: Replicase polyprotein 1a (pp1a) (ORF1a)] [Contains: p9; p87; p195 (Papain-like proteinases 1/2) (PL1-PRO/PL2-PRO); Peptide HD2; 3C-like proteinase (3CL-PRO) (3CLp) (M-PRO) (p34); Unknown protein 1; p5; p23; p12; Growth factor-like peptide (GFL) (p16); RNA-directed RNA polymerase (RdRp) (Pol) (p100); Helicase (Hel) (p66) (p66-HEL); Unknown protein 2; p41; Unknown protein 3]
gi|12082740|gb|AAG48591.1| replicase polyprotein 1ab [Human coronavirus 229E]
Length = 6758

Score = 429 bits (1104), Expect = e-119
Identities = 233/535 (43%), Positives = 323/535 (60%), Gaps = 18/535 (3%)
Frame = +2

```
Query: 41    IERYLENSSENGIPLMPLLSCGIFGVRIENSLKALFSCDINKPLQVFVYSSNEEQAVLKF 220
             I+ Y    ++E G PL P+LSCGIFG+++E SL+ L     K ++VFVY+  E   V  F
Sbjct: 1372 IKAYNTINNEQGTPLTPILSCGIFGIKLETSLEVLLDVCNTKEVKVFVYTDTEVCKVKDF 1431

Query: 221   LDGLDLTPVIDD-------VDVVK----PFRVEGNFSFFDCGVNAL-DGDIYLLFTNSIL 364
             + GL      ++       V V+K    P+RV+G FS+F    +  + D    +LFT+S+L
Sbjct: 1432 VSGLVNVQKVEQPKIEPKPVSVIKVAPKPYRVDGKFSYFTEDLLCVADDKPIVLFTDSML 1491

Query: 365   MLDKQGQLLDTKLNGILQQAVLDYLATVKTVPAGNLVKLVVESCTIYMCVVPSINDLSFD 544
              LD +G  LD  L+G+L  A+ D +    K +P+GNL+K  + S  +YMCVVPS   D  D
Sbjct: 1492 TLDDRGLALDNALSGVLSAAIKDCVDINKAIPSGNLIKFDIGSVVVYMCVVPSEKDKHLD 1551

Query: 545   KNLGRCVRKLNRLKTCVIANVPAIDXXXXXXXXXXXXXXFVVESNVMDVNDCFKNDNVVL 724
              N+ RC RKLNRL   ++  +PA                FV E     +         + +
Sbjct: 1552 NNVQRCTRKLNRLMCDIVCTIPADYILPLVLSSLTCNVSFVGELKAAEAKV------ITI 1605

Query: 725   KITEDGINVKDVVVESSKSLGKQLGVVSDGVDSFEGVLP--INTDTVLSVAPEVDWVAFY 898
             K+TEDG+NV DV V + KS  +Q+GV++D        G +P  +NT  +L+ A +VDWV FY
Sbjct: 1606 KVTEDGVNVHDVTVTTDKSFEQQVGVIADKDKDLSGAVPSDLNTSELLTKAIDVDWVEFY 1665

Query: 899   GFEKAALFASLDVKPYGYPNDFVGGFRVLGTTDNNCWVNATCIILQYLKPTFKSKGLNVL 1078
             GF+ A  FA++D   + Y +  V G RVL T+DNNCWVNA CI LQY KP F S+GL+
Sbjct: 1666 GFKDAVTFATVDHSAFAYESAVVNGIRVLKTSDNNCWVNAVCIALQYSKPHFISQGLDAA 1725

Query: 1079  WNKFVTGDVGPFVSFIYFITMSSKGQKGDAEEALSKLSEYLISDSIVTLEQYSTCDIC-- 1252
             WNKFV GDV  FV+F+Y++      KG KGDAE+ L+KLS+YL +++ V LE YS+C  C
Sbjct: 1726 WNKFVLGDVEIFVAFVYYVARLMKGDKGDAEDTLTKLSKYLANEAQVQLEHYSSCVECDA 1785

Query: 1253  --KSTVVEVKSAVVCASVLKDGCDVGFCPHRHKLRSRVKFVNGRVVITNVGEPIISQPSK 1426
               K++V  + SA+VCASV +DG  VG+C H  K  SRV+ V GR +I +V +      S+
Sbjct: 1786 KFKNSVASINSAIVCASVKRDGVQVGYCVHGIKYYSRVRSVRGRAIIVSVEQLEPCAQSR 1845

Query: 1427  LLNGIAYTTFSGSFDNGHYVVYDAANNAVYDGARLFASDLSTLAVTAIVVVGGCV 1591
             LL+G+AYT FSG  D GHY VYD A  ++YDG R    DLS L+VT++V+VGG V
Sbjct: 1846 LLSGVAYTAFSGPVDKGHYTVYDTAKKSMYDGDRFVKHDLSLLSVTSVVMVGGYV 1900
```

Fig 3 (cont)

3. Sequence C

6017 nucleotides; Encodes part of Replicase

CGAGAACAGCTTGATTCGTTATTTGTTATACTTGTATTTTGTAATTTTTGGTAAACGTATTTGCGTTTTGGACTT
TTATATTTTGTTGCACAATTTATTAGTACTTTTGGTTCTTTCTTAGGCTTTCATCAGAAACAGTGGTTTTTACAT
TTTGTGCCGTTTGATGTTTTATGTAATGAGTTTTtAGCTaCATTTATTGTCTGCAAAATTGTTTTATTTGTTAGA
CATATTATTGTTGGCTGTAATAATGCTGACTGTGTAGCTTGTTCTAAAAGTGCTAGACTTAAACGTGTACCACTT
CAAACTATTATTAATGGTATGCATAAATCATTCTATGTTAATGCTAATGGTGGTACTTGTTTCTGTAATAAACAT
AACTTCTTTTGTGTTAATTGTGATTCTTTTGGGCCTGGTAATACTTTTATTAATGGTGATATTGCAAGAGAGCTT
GGTAATGTTGTTAAAACAGCTGTTCAACCCACAGCTCCTGCATATGTTATTATTGATAAGGTAGATTTTGTTAAT
GGATTTTATCGTCTTTATAGTGGTGACACTTTTTGGCGGTATGACTTTGACATTACTGAATCTAAGTATAGTTGT
AAAGAGGTTCTGAAGAATTGTAATGTTTTAGAAAATTTTATTGTTTACAATAATAGTGGTAGTAACATTACACAG
ATTAAAAATGCTTGTGTTTATTTTTCTCAATTGTTGTGTGAACCTATAAAGTTGGTAAATTCAGAGTTGTTGTCA
ACTTTATCAGTTGATTTTAATGGTGTTTTGCATAAGGCATATGTTGATGTTTTGTGTAATAGTTTTTTTAAGGAG
CTAACTGCTAACATGTCCATGGCTGAATGTAAAGCTACACTTGGTTTGACTGTTTCTGATGATGATTTTGTTTCA
GCTGTTGCCAATGCACATAGGTATGACGTTTTGCTTTCAGATTTGTCATTTAATAATTTTTTTATTTCTTATGCT
AAACCTGAAGATAAGTTGTCCGTTTATGACATTGCTTGTTGTATGCGTGCCGGTTCTAAGGTTGTTAACCATAAT
GTTTTAATCAAAGAGTCAATACCTATTGTTTGGGGTGTCAAGGACTTTAATACTCTTTCTCAAGAAGGTAAGAAG
TACCTTGTTAAAACAACTAAAGCAAAGGGTTTGACTTTTTTATTAACTTTTAATGATAACCAAGCAATTACACAA
GTTCCTGCTACTAGTATAGTTGCAAAACAGGGTGCTGGTTTTAAACGTACTTATAATTTTCTGTGGTATGTATGT
TTATTTGTTGTTGCATTGTTTATTGGTGTCTCATTTATTGATTATACAACCACTGTAACTAGCTTTCATGGTTAT
GATTTTAAGTACATTGAGAATGGTCAGTTGAAGGTGTTTGAAGCACCTTTACACTGTGTTCGTAATGTTTTTGAT
AATTTTAATCAATGGCATGAGGCTAAGTTTGGTGTTGTTACTACTAATAGTGATAAATGTCCTATAGTTGTTGGT
GTTTCAGAGCGTATTAATGTTGTTCCTGGTGTTCCAACAAATGTATATTTGGTAGGAAAGACTCTTGTTTTTACA
TTACAGGCTGCTTTTGGAAACACAGGTGTTTGTTATGACTTTGATGGTGTTACCACTAGTGATAAGTGTATTTTT
AATTCTGCTTGTACTAGGTTGGAAGGTTTGGGTGGTGACAATGTTTATTGTTACAACACTGATCTTATTGAAGGT
TCTAAACCTTATAGTATTTTACAGCCCAATGCTTATTATAAGTATGATGTTAAAAATTATGTACGTTTTCCAGAA
ATTTTAGCTAGAGGTTTTGGCTTACGTACTATTAGAACTTTGGCTACACGTTATTGTAGAGTTGGTGAATGCCGT
GACTCACATAAAGGTGTTTGTTTTGGTTTTGATAAATGGTATGTTAATGATGGACGTGTTGATGACGGTTACATT
TGTGGTGATGGTCTTATAGACCTTCTTGTTAATGTACTCTCAATCTTTAGTTCATCTTTTAGCGTTGTGGCTATG
TCTGGACATATGTTGTTTAATTTTCTTTTTGCAKCATTTATTACATTTTTGTGCTTTTTAGTTACTAAATTTAAA
CGTGTTTTTGGTGATCTTTCTTATGGTGTTTTTACTGTTGTTTGTGCAACTTTGATTAATAACATTTCTTATGTT
GTTACTCAAAATTTATTTTTTATGTTGCTTTATGCTATTTTGTATTTTGTTTTTACTAGGACAGTGCGTTATGCT
TGGATTTGGCATATTGCATACATTGTTGCATACTTCTTGTTAATACCATGGTGGCTTCTCACATGGTTTAGTTTT
GCTGCATTTTTAGAGCTTTTACCTAATGTTTTTAAGTTAAAAATCTCTACTCAATTGTTTGAAGGTGATAAGTTT
ATAGGTACTTTTGAGAGTGCTGCTGCAGGTACATTTGTTCTTGACATGCGTTCTTATGAAAGGCTGATAAATACT
ATTTCACCTGAGAAACTTAAGAATTATGCTGCAAGTTATAATAAATATAAATATTATAGTGGTAGTGCTAGTGAG
GCTGATTATCGTTGTGCTTGTTATGCTCATTTAGCCAAGGCTATGTTAGATTACGCAAAAGATCATAATGACATG
TTATATTCTCCACCTACCATTAGCTACAATTCCACCTTACAATCTGGTCTTAAGAAGATGGCACAACCATCTGGT
TGTGTTGAGAGATGTGTGGTTCGCGTCTGTTATGGTAGTACTGTGCTTAATGGAGTTTGGTTAGGTGACACTGTT
ACTTGTCCTAGACATGTCATAGCACCATCAACCACTGTTCTTATTGATTATGATCATGCATATAGTACTATGCGT
TTGCATAATTTTTCAGTGTCTCATAATGGTGTCTTCTTGGGAGTTGTTGGTGTTACAATGCATGGTTCTGTGTTG
CGTATTAAGGTTTCACAATCTAATGTACATACACCTAAACATGTTTTTAAAACGTTGAAACCTGGTGCTTCTTTT
AATATTTTAGCATGTTATGAAGGTATTGCATCTGGTGTTTTTGGTGTTAATTTACGTACAAACTTtACTAtTAAA
GGTTCTTTTAtAAATGGAGCTTGTGGTTCTCCTGGTTATAATGTTAGAAATGATGGTACTGTTGAGTTTTGTTAT
TTACACCAAATTGAGTTAGGTAGTGGTGCTCATGTTGGTTCTGATTTTACTGGTAGTGTTTATGGTAATTTTGAT
GACCAACCTAGTTTGCAAGTTGAGAGTGCCAACCTTATGCTATCAGATAATGTTGTTGCCTTTTTGTATGCTGCT
TTGTTGAATGGTTGTAGGTGGTGGTTGCGTTCAACTAGAGTTAATGTTGATGGTTTTAATGAATGGGCTATGGCT
AATGGTTATACAATTGTTTCTAGTGTTGAGTGCTATTCTATTTTGGCAGCAAAAACTGGTGTTAGTGTTGAACAA
TTGTTAGCTTCCATTCAACATCTTCATGAAGGTTTTGGTGGTAAAAACATACTTGGTTATTCTAGTTTATGTGAT
GAGTTCACACTAGCTGAAGTTGTGAAGCAGATGTATGGTGTTAACTTGCAAAGTGGTAAGGTTATTTTTGGTTTA
AAAACAATGTTTTTATTTAGCGTTTTCTTCACAATGTTTTGGGCAGAACTCTTTATTTATACAAACACTATATGG
ATAAACCCTGTTATACTTACACCTATATTTGTTTACTTTTGTTTTTGTCATTAGTTTTAACTATGTTTCTTAAA
CATAAGTTTTTGTTTTTGCAAGTATTTTTATTACCTACTGTTATTGCAACTGCTTTATATAATTGTGTTTTGGAT
TATTACATAGTAAAATTTTTGGCTGACCATTTTAACTATAATGTTTCAGTATTACAAATGGATGTTCAGGGTTTA
GTTAATGTTTTGGTCTGTTTATTTGTTGTATTTTTACACACATGGCGTTTTTCTAAAGAACGTTTCACACATTGG
TTTACATATGTGTGTTCTCTTATAGCAGTTGCTTACACTTATTTTTATAGTGGTGACTTTTTGAGTTTGCTTGTT
ATGTTTTTATGTGCTATATCTAGTGATTGGTACATTGGTGCCATTGTTTTTAGGTTGTCACGTTTGATTATATTT
TTTTCACCTGAAAGTGTATTTAGTGTTTTTGGTGATGTGAAACTCACTTTAGTTGTTTATTTAATTTGTGGTTAT
TTAGTTTGTACTTATTGGGGCATTTTGTATTGGTTCAaTAGGTTTTTTAAATGTACTATGGGTGTTTATGATTTT
AAGGTGAGTGCTGCTGAATTTAAATACATGGTTGCTAATGGACTTCATGCACCATATGGACCTTTTGATGCACTT
TGGTTATCATTCAAATTACTTGGTATTGGTGGTGACCGTTGTATAAAAATTTCAACTGTCCAATCCAAACTGACT
GATTTGAAGTGTACTAATGTTGTGTTATTGGGTTGTTTGTCTAGTATGAACATTGCAGCTAATTCTAGTGAATGG
GCTTATTGTGTTGATTTACACAATAAGATTAATCTTTGTGATGACCCAGAAAAAGCTCAAGGTATGTTGTTAGCA
CTCCTTGCGTTCTTTCTAAGTAAACATAGTGATTTTGGTCTTGATGGCCTTATTGATTCTTATTTTGATAATAGT
AGCACCCTGCAGAGTGTTGCTTCATCATTTGTTAGTATGCCATCATATATTGCTTATGAAAATGCTAGACAAGCT
TATGAGGATGCTATTGCTAATGGATCTTCTTCTCAACTTATTAAACAATTGAAGCGTGCCATGAATATCGCAAAG
TCTGAATTTGATCATGAGATATCTGTTCAGAAGAAAATTAATAGAATGGCTGAACAAGCTGCTACTCAGATGTAT
AAAGAAGCACGCTCTGTTAATAGAAAATCTAAGTTATTAGTGCTATGCACTCTTTACTTTTTGGAATGTTAAGA

CGTTTGGATATGTCTAGTGTTGAAACTGTTTTGAATTTAGCACGTGATGGTGTTGTGCCATTGTCAGTTATACC
GCAACTTCAGCTTCCAAACTAACTATTGTTAGTCCAGATCTTGAATCTTATTCTAAGATTGTTTGTGATGGTTC
GTTCATTATGCTGGAGTTGTTTGGACACTTAATGATGTTAAAGACAATGATGGTAGACCTGTTCATGTTAAAGA
ATTACAAGGGAGAATGTTGAAACTTTGACATGGCCTCTTATCCTTAATTGTGAACGTGTTGTTAAACTTCAAAA
AATGAAATTATGCCTGGTAAACTTAAGCAAAAACCTATGAAAGCTGAGGGTGATGGTGGTGTTTTAGGTGATGG
AATGCTTTGTATAATACTGAGGGTGGTAAAACTTTTATGTATGCTTATATTTCTAATAAAGCTGACCTTAAATT
GTTAAGTGGGAGTATGAGGGTGGTTGCAACACAATCGAGTTAGACTCTCCTTGTCGATTTATGGTCGAAACACC
AATGGTCCTCAAGTGAAGTATTTGTATTTTGTTAAAAATTTAAATACCTTACGTAGAGGTGCCGTTCTTGGTTT
ATAGGTGCCACAATTCGTCTACAAGCTGGTAAACAAACTGAATTGGCTGTTAATTCTGGACTTTTAACTGCTTG
GCTTTTTCTGTTGATCCAGCAACCACTTACTTGGAAGCTGTTAAACATGGTGCAAAACCTGTAAGTAATTGTAT
AAGATGTTATCTAATGGTGCTGGTAATGGTCAAGCTATAACAACTAGTGTAGATGCTAACACCAATCAAGATTC
TATGGTGGAGCGTCTATTTGTTTGTATTGTCGGGCCCACGTTCCTCACCCTAGTATGGATGGTTACTGTAAGTT
AAGGGTAAATGTGTTCAGGTTCCTATTGGTTGTTTGGATCCTATTAGGTTTTGTTTAGAAAATAATGTGTGTAA
GTTTGTGGTTGTTGGTTGGGACACGGGTGTGCTTGTGATCGTACAACCATTCAAAGTGTTGACATTCTTATTTA
ACGAACGATCAAGCTGT

Putative ORFs

>~out: 55 to 5997: Frame 1 1981 aa
TYLRFGLLYFVAQFISTFGSFLGFHQKQWFLHFVPFDVLCNEFLATFIVCKIVLFVRHIIVGCNNADCVACSKS
RLKRVPLQTIINGMHKSFYVNANGGTCFCNKHNFFCVNCDSFGPGNTFINGDIARELGNVVKTAVQPTAPAYVI
DKVDFVNGFYRLYSGDTFWRYDFDITESKYSCKEVLKNCNVLENFIVYNNSGSNITQIKNACVYFSQLLCEPIK
VNSELLSTLSVDFNGVLHKAYVDVLCNSFFKELTANMSMAECKATLGLTVSDDDFVSAVANAHRYDVLLSDLSF
NFFISYAKPEDKLSVYDIACCMRAGSKVVNHNVLIKESIPIVWGVKDFNTLSQEGKKYLVKTTKAKGLTFLLTF
DNQAITQVPATSIVAKQGAGFKRTYNFLWYVCLFVVALFIGVSFIDYTTTVTSFHGYDFKYIENGQLKVFEAPL
CVRNVFDNFNQWHEAKFGVVTTNSDKCPIVVGVSERINVVPGVPTNVYLVGKTLVFTLQAAFGNTGVCYDFDGV
TSDKCIFNSACTRLEGLGGDNVYCYNTDLIEGSKPYSILQPNAYYKYDVKNYVRFPEILARGFGLRTIRTLATR
CRVGECRDSHKGVCFGFDKWYVNDGRVDDGYICGDGLIDLLVNVLSIFSSSFSVVAMSGHMLFNFLFAXFITFL
FLVTKFKRVFGDLSYGVFTVVCATLINNISYVVTQNLFFMLLYAILYFVFTRTVRYAWIWHIAYIVAYFLLIPW
LLTWFSFAAFLELLPNVFKLKISTQLFEGDKFIGTFESAAAGTFVLDMRSYERLINTISPEKLKNYAASYNKYK
YSGSASEADYRCACYAHLAKAMLDYAKDHNDMLYSPPTISYNSTLQSGLKKMAQPSGCVERCVVRVCYGSTVLN
VWLGDTVTCPRHVIAPSTTVLIDYDHAYSTMRLHNFSVSHNGVFLGVVGVTMHGSVLRIKVSQSNVHTPKHVFK
LKPGASFNILACYEGIASGVFGVNLRTNFTIKGSFINGACGSPGYNVRNDGTVEFCYLHQIELGSGAHVGSDFT
SVYGNFDDQPSLQVESANLMLSDNVVAFLYAALLNGCRWWLRSTRVNVDGFNEWAMANGYTIVSSVECYSILAA
TGVSVEQLLASIQHLHEGFGGKNILGYSSLCDEFTLAEVVKQMYGVNLQSGKVIFGLKTMFLFSVFFTMFWAEL
IYTNTIWINPVILTPIFCLLLFLSLVLTMFLKHKFLFLQVFLLPTVIATALYNCVLDYYIVKFLADHFNYNVSV
QMDVQGLVNVLVCLFVVFLHTWRFSKERFTHWFTYVCSLIAVAYTYFYSGDFLSLLVMFLCAISSDWYIGAIVF
LSRLIIFFSPESVFSVFGDVKLTLVVYLICGYLVCTYWGILYWFNRFFKCTMGVYDFKVSAAEFKYMVANGLHA
YGPFDALWLSFKLLGIGGDRCIKISTVQSKLTDLKCTNVVLLGCLSSMNIAANSSEWAYCVDLHNKINLCDDPE
AQGMLLALLAFFLSKHSDFGLDGLIDSYFDNSSTLQSVASSFVSMPSYIAYENARQAYEDAIANGSSSQLIKQL
RAMNIAKSEFDHEISVQKKINRMAEQAATQMYKEARSVNRKSKVISAMHSLLFGMLRRLDMSSVETVLNLARDG
VPLSVIPATSASKLTIVSPDLESYSKIVCDGSVHYAGVVWTLNDVKDNDGRPVHVKEITRENVETLTWPLILNC
RVVKLQNNEIMPGKLKQKPMKAEGDGGVLGDGNALYNTEGGKTFMYAYISNKADLKFVKWEYEGGCNTIELDSPC
RFMVETPNGPQVKYLYFVKNLNTLRRGAVLGFIGATIRLQAGKQTELAVNSGLLTACAFSVDPATTYLEAVKHGA
KPVSNCIKMLSNGAGNGQAITTSVDANTNQDSYGGASICLYCRAHVPHPSMDGYCKFKGKCVQVPIGCLDPIRFC
LENNVCNVCGCWLGHGCACDRTTIQSVDILI
>~out: 263 to 511: Frame 2 83 aa
LVLKVLDLNVYHFKLLLMVCINHSMLMLMVVLVSVINITSFVLIVILLGLVILLLMVILQESLVMLLKQLFNPQL
LHMLLLIR
>~out: 875 to 1054: Frame 2 60 aa
LFLMMILFQLLPMHIGMTFCFQICHLIIFLFLMLNLKISCPFMTLLVVCVPVLRLLTIMF
>~out: 1556 to 1804: Frame 2 83 aa
ERLLFLHYRLLLETQVFVMTLMVLPLVISVFLILLVLGWKVWVVTMFIVTTLILLKVLNLIVFYSPMLIISMMLK
IMYVFQKF
>~out: 1808 to 1966: Frame 2 53 aa
LEVLAYVLLELWLHVIVELVNAVTHIKVFVLVLINGMLMMDVLMTVTFVVMVL
>~out: 2600 to 2761: Frame 2 54 aa
ITQKIIMTCYILHLPLATIPPYNLVLRRWHNHLVVLRDVWFASVMVVLCLMEFG
>~out: 2798 to 2980: Frame 2 61 aa
HHQPLFLLIMIMHIVLCVCIIFQCLIMVSSWELLVLQCMVLCCVLRFHNLMYIHLNMFLKR
>~out: 4595 to 4774: Frame 2 60 aa
VNIVILVLMALLILILIIVAPCRVLLHHLLVCHHILLMKMLDKLMRMLLLMDLLLNLLNN
>~out: 4790 to 4945: Frame 2 52 aa
ISQSLNLIMRYLFRRKLIEWLNKLLLRCIKKHALLIENLKLLVLCTLYFLEC
>~out: 5048 to 5200: Frame 2 51 aa
LLLVQILNLILRLFVMVLFIMLELFGHLMMLKTMMVDLFMLKRLQGRMLKL

Fig 3. (cont) 12/25

```
>~out: 5753 to 5905: Frame 2       51 aa
MLTPIKILMVERLFVCIVGPTFLTLVWMVTVSLRVNVFRFLLVVWILLGFV
```

Alignment

>gi|12175747|ref|NP_073549.1| replicase polyprotein 1ab [Human coronavirus 229E]
gi|30179827|sp|Q05002|R1AB_CVH22 Replicase polyprotein 1ab (pp1ab) (ORF1ab polyprotein) [Includes: Replicase polyprotein 1a (pp1a) (ORF1a)] [Contains: p9; p87; p195 (Papain-like proteinases 1/2) (PL1-PRO/PL2-PRO); Peptide HD2; 3C-like proteinase (3CL-PRO) (3CLp) (M-PRO) (p34); Unknown protein 1; p5; p23; p12; Growth factor-like peptide (GFL) (p16); RNA-directed RNA polymerase (RdRp) (Pol) (p100); Helicase (Hel) (p66) (p66-HEL); Unknown protein 2; p41; Unknown protein 3]
gi|12082740|gb|AAG48591.1| replicase polyprotein 1ab [Human coronavirus 229E]
Length = 6758

Score = 2840 bits (7361), Expect = 0.0
Identities = 1350/1997 (67%), Positives = 1609/1997 (80%), Gaps = 4/1997 (0%)
Frame = +1

```
Query: 10    LDSLFVILVFCNFW*TYLRFGLLYFVAQFISTFGSFLGFHQKQWFLHFVPFDVLCNEFLA 189
              +    V+++    F    YLR  LLYFVAQ IST G FLG+ +   WFLHF+PFDV+C+E L
Sbjct: 2076  MQPFIVMVLLLIFGDNYLRCFLLYFVAQMISTVGVFLGYKETNWFLHFIPFDVICDELLV 2135

Query: 190   TFIVCKIVLFVRHIIVGCNNADCVACSKSARLKRVPLQTIINGMHKSFYVNANGGTCFCN 369
              T IV K++ FVRH++ GC N DC+ACSKSARLKR P+ TI+NG+ +SFYVNANGG+ FC
Sbjct: 2136  TVIVIKVISFVRHVLFGCENPDCIACSKSARLKRFPVNTIVNGVQRSFYVNANGGSKFCK 2195

Query: 370   KHNFFCVNCDSFGPGNTFINGDIARELGNVVKTAVQPTAPAYVIIDKVDFVNGFYRLYSG 549
              KH FFCV+CDS+G G+TFI   +++RELGN+ KT VQPT PAYV+IDKV+F NGFYRLYS
Sbjct: 2196  KHRFFCVDCDSYGYGSTFITPEVSRELGNITKTNVQPTGPAYVMIDKVEFENGFYRLYSC 2255

Query: 550   DTFWRYDFDITESKYSCKEVLKNCNVLENFIVYNNSGSNITQIKNACVYFSQLLCEPIKL 729
              +TFWRY+FDITESKYSCKEV KNCNVL++FIV+NN+G+N+TQ+KNA VYFSQLLC PIKL
Sbjct: 2256  ETFWRYNFDITESKYSCKEVFKNCNVLDDFIVFNNNGTNVTQVKNASVYFSQLLCRPIKL 2315

Query: 730   VNSELLSTLSVDFNGVLHKAYVDVLCNSFFKELTANMSMAECKATLGLTVSDDDFVSAVA 909
              V+SELLSTLSVDFNGVLHKAY+DVL NSF K+L ANMS+AECK  LGL++SD +F SA++
Sbjct: 2316  VDSELLSTLSVDFNGVLHKAYIDVLRNSFGKDLNANMSLAECKRALGLSISDHEFTSAIS 2375

Query: 910   NAHRYDVLLSDLSFNNFFISYAKPEDKLSVYDIACCMRAGSKVVNHNVLIKESIPIVWGV 1089
              NAHR DVLLSDLSFNNF  SYAKPE+KLS YD+ACCMRAG+KVVN NVL K+   PIVW
Sbjct: 2376  NAHRCDVLLSDLSFNNFVSSYAKPEEKLSAYDLACCMRAGAKVVNANVLTKDQTPIVWHA 2435

Query: 1090  KDFNTLSQEGKKYLVKTTKAKGLTFLLTFNDNQAITQVPATSIVAKQGAGFK-RTYNFLW 1266
              KDFN+LS EG+KY+VKT+KAKGLTFLLT N+NQA+TQ+PATSIVAKQGAG    +  +LW
Sbjct: 2436  KDFNSLSAEGRKYIVKTSKAKGLTFLLTINENQAVTQIPATSIVAKQGAGDAGHSLTWLW 2495

Query: 1267  YVCLFVVAL-FIGVSFIDYTT--TVTSFHGYDFKYIENGQLKVFEAPLHCVRNVFDNFNQ 1437
               +C  V  + F    F+ Y     V+SF GYDFKYIENGQLK FEAPL CVRNVF+NF
Sbjct: 2496  LLCGLVCLIQFYLCFFMPYFMYDIVSSFEGYDFKYIENGQLKNFEAPLKCVRNVFENFED 2555

Query: 1438  WHEAKFGVVTTNSDKCPIVVGVSERINVVPGVPTNVYLVGKTLVFTLQAAFGNTGVCYDF 1617
              WH AKFG    N    CPIVVGVSE +N V  G+P+NVYLVGKTL+FTLQAAFGN GVCYD
Sbjct: 2556  WHYAKFGFTPLNKQSCPIVVGVSEIVNTVAGIPSNVYLVGKTLIFTLQAAFGNAGVCYDI 2615

Query: 1618  DGVTTSDKCIFNSACTRLEGLGGDNVYCYNTDLIEGSKPYSILQPNAYYKYDVKNYVRFP 1797
               GVTT +KCIF SACTRLEGLGG+NVYCYNT L+EGS PYS +Q NAYYKYD  N+++ P
Sbjct: 2616  FGVTTPEKCIFTSACTRLEGLGGNNVYCYNTALMEGSLPYSSIQANAYYKYDNGNFIKLP 2675

Query: 1798  EILARGFGLRTIRTLATRYCRVGECRDSHKGVCFGFDKWYVNDGRVDDGYICGDGLIDXX 1977
              E++A+GFG RT+RT+AT+YCRVGEC +S+  GVCFGFDKW+VNDGRV +GY+CG GL +
Sbjct: 2676  EVIAQGFGFRTVRTIATKYCRVGECVESNAGVCFGFDKWFVNDGRVANGYVCGTGLWNLV 2735

Query: 1978  XXXXXXXXXXXXXXAMSGHMLFNFLFAXFITFLCFLVTKFKRVFGDLSYGVFTVVCATLI 2157
                            AMSG +L N       F  F CFLVTKF+R+FGDLS GV TVV A L+
Sbjct: 2736  FNILSMFSSSFSVAAMSGQILLNCALGAFAIFCCFLVTKFRRMFGDLSVGVCTVVVAVLL 2795

Query: 2158  NNISYVVTQNLFFMLLYAILYFVFTRTVRYAWIWHIAYIVAYFLLIPWWLLTWFSFAAFL 2337
              NN+SY+VTQNL   M+ YAILYF   TR++RYAWIW  AY++AY    PWWL  W+  A
```

Fig 3 (Cont) 13/25

```
Sbjct: 2796 NNVSYIVTQNLVTMIAYAILYFFATRSLRYAWIWCAAYLIAYISFAPWWLCAWYFLAMLT 2855

Query: 2338 ELLPNVFKLKISTQLFEGDKFIGTFESAAAGTFVLDMRSYERLINTISPEKLXXXXXXXX 2517
             LLP++ KLK+ST LFEGDKF+GTFESAAAGTFV+DMRSYE+L N+ISPEKL
Sbjct: 2856 GLLPSLLKLKVSTNLFEGDKFVGTFESAAAGTFVIDMRSYEKLANSISPEKLKSYAASYN 2915

Query: 2518 XXXXXXXXXXEADYRCACYAHLAKAMLDYAKDHNDMLYSPPTISYNSTLQSGLKKMAQPS 2697
                      EADYRCACYA+LAKAMLD+++DHND+LY+PPT+SY STLQ+GL+KMAQPS
Sbjct: 2916 RYKYYSGNANEADYRCACYAYLAKAMLDFSRDHNDILYTPPTVSYGSTLQAGLRKMAQPS 2975

Query: 2698 GCVERCVVRVCYGSTVLNGVWLGDTVTCPRHVIAPSTTVLIDYDHAYSTMRLHNFSVSHN 2877
            G VE+CVVRVCYG+TVLNG+WLGD V CPRHVIA +TT  IDYDH YS MRLHNFS+
Sbjct: 2976 GFVEKCVVRVCYGNTVLNGLWLGDIVYCPRHVIASNTTSAIDYDHEYSIMRLHNFSIISG 3035

Query: 2878 GVFLGVVGVTMHGSVLRIKVSQSNVHTPKHVFKTLKPGASFNILACYEGIASGVFGVNLR 3057
              FLGVVG TMHG  L+IKVSQ+N+HTP+H F+TLK G  FNILACY+G A GVFGVN+R
Sbjct: 3036 TAFLGVVGATMHGVTLKIKVSQTNMHTPRHSFRTLKSGEGFNILACYDGCAQGVFGVNMR 3095

Query: 3058 TNFTIKGSFINGACGSPGYNVRNDGTVEFCYLHQIELGSGAHVGSDFTGSVYGNFDDQPS 3237
            TN+TI+GSFINGACGSPGYN++N G VEF Y+HQIELGSG+HVGS F G +YG F+DQP+
Sbjct: 3096 TNWTIRGSFINGACGSPGYNLKN-GEVEFVYMHQIELGSGSHVGSSFDGVMYGGFEDQPN 3154

Query: 3238 LQVESANLMLSDNVVAFLYAALLNGCRWWLRSTRVNVDGFNEWAMANGYTIVSSVECYSI 3417
            LQVESAN ML+ NVVAFLYAA+LNGC WWL+  ++ V+ +NEWA ANG+T ++   + +SI
Sbjct: 3155 LQVESANQMLTVNVVAFLYAAILNGCTWWLKGEKLFVEHYNEWAQANGFTAMNGEDAFSI 3214

Query: 3418 LAAKTGVSVEQLLASIQHLHEGFGGKNILGYSSLCDEFTLAEVVKQMYGVNLQSGKVIFG 3597
            LAAKTGV VE+LL +IQ L+ GFGGK ILGYSSL DEF++ EVVKQM+GVNLQSGK
Sbjct: 3215 LAAKTGVCVERLLHAIQVLNNGFGGKQILGYSSLNDEFSINEVVKQMFGVNLQSGKTTSM 3274

Query: 3598 LKTMFLFSVFFTMFWAELFIYTNTIWINPVIXXXXXXXXXXXXXXXXXXXKHKFLFLQVF 3777
             K++ LF+ FF MFWAELF+YT TIW+NP                     KHK LFLQVF
Sbjct: 3275 FKSISLFAGFFVMFWAELFVYTTTIWVNPGFLTPFMILLVALSLCLTFVVKHKVLFLQVF 3334

Query: 3778 LLPTVIATALYNCVLDYYIVKFLADHFNYNVSVLQMDVQGXXXXXXXXXXXXXHTWRFSK 3957
            LLP++I  A+ NC  DY++ K LA+ F+YNVSV+QMD+QG             HTWRF+K
Sbjct: 3335 LLPSIIVAAIQNCAWDYHVTKVLAEKFDYNVSVMQMDIQGFVNIFICLFVALLHTWRFAK 3394

Query: 3958 ERFTHWFTYVCSLIAVAYTYFYSGDFLSLLVMFLCAISSDWYIGAIVFRLSRLIIFFSPE 4137
            ER THW TY+ SLIAV YT  YS D++SLLVM LCAIS++WYIGAI+FR+ R  + F P
Sbjct: 3395 ERCTHWCTYLFSLIAVLYTALYSYDYVSLLVMLLCAISNEWYIGAIIFRICRFGVAFLPV 3454

Query: 4138 SVFSVFGDVKLTLVVYLICGYLVCTYWGILYWFNRFFKCTMGVYDFKVSAAEFKYMVANG 4317
               S F  VK  L+ Y++ G++ C Y+G+LYW NRF KCT+GVYDF VS AEFKYMVANG
Sbjct: 3455 EYVSYFDGVKTVLLFYMLLGFVSCMYYGLLYWINRFCKCTLGVYDFCVSPAEFKYMVANG 3514

Query: 4318 LHAPYGPFDALWLSFKLLGIGGDRCIKISTVQSKLTDLKCTNVVLLGCLSSMNIAANSSE 4497
            L+AP GPFDAL+LSFKL+GIGG R IK+STVQSKLTDLKCTNVVL+G LS+MNIA+NS E
Sbjct: 3515 LNAPNGPFDALFLSFKLMGIGGPRTIKVSTVQSKLTDLKCTNVVLMGILSNMNIASNSKE 3574

Query: 4498 WAYCVDLHNKINLCDDPEKAQGMLLALLAFFLSKHSDFGLDGLIDSYFDNSSTLQSVASS 4677
            WAYCV++HNKINLCDDPE AQ +LLALLAFFLSKHSDFGL  L+DSYF+N S LQSVASS
Sbjct: 3575 WAYCVEMHNKINLCDDPETAQELLLALLAFFLSKHSDFGLGDLVDSYFENDSILQSVASS 3634

Query: 4678 FVSMPSYIAYENARQAYEDAIANGSSSQLIKQLKRAMNIAKSEFDHEISVQKKINRMAEQ 4857
            FV MPS++AYE ARQ YE+A+ANGSS Q+IKQLK+AMN+AK+EFD E SVQKKINRMAEQ
Sbjct: 3635 FVGMPSFVAYETARQEYENAVANGSSPQIIKQLKKAMNVAKAEFDRESSVQKKINRMAEQ 3694

Query: 4858 AATQMYKEARSVNRKSKVISAMHSLLFGMLRRLDMSSVETVLNLARDGVVPLSVIPATSA 5037
            AA  MYKEAR+VNRKSKV+SAMHSLLFGMLRRLDMSSV+T+LN+AR+GVVPLSVIPATSA
Sbjct: 3695 AAAAMYKEARAVNRKSKVVSAMHSLLFGMLRRLDMSSVDTILNMARNGVVPLSVIPATSA 3754

Query: 5038 SKLTIVSPDLESYSKIVCDGSVHYAGVVWTLNDVKDNDGRPVHVKEITRENVETLTWPLI 5217
            ++L +V PD +S+ K++ DG VHYAGVVWTL +VKDNDG+ VH+K++T+EN E L WPLI
Sbjct: 3755 ARLVVVVPDHDSFVKMMVDGFVHYAGVVWTLQEVKDNDGKNVHLKDVTKENQEILVWPLI 3814

Query: 5218 LNCERVVKLQNNEIMPGKLKQKPMKAEGDGGVLGDGNALYNTEGGKTFMYAYISNKADLK 5397
            L CERVVKLQNNEIMPGK+K K  K EGDGG+  +GNALYN EGG+ FMYAY++ K  +K
Sbjct: 3815 LTCERVVKLQNNEIMPGKMKVKATKGEGDGGITSEGNALYNNEGGRAFMYAYVTTKPGMK 3874

Query: 5398 FVKWEYEGGCNTIELDSPCRFMVETPNGPQVKYLYFVKNLNTLRRGAVLGFIGATIRLQA 5577
            +VKWE++ G  T+EL+ PCRF+++TP GPQ+KYLYFVKNLN LRRGAVLG+IGAT+RLQA
Sbjct: 3875 YVKWEHDSGVVTVELEPPCRFVIDTPTGPQIKYLYFVKNLNNLRRGAVLGYIGATVRLQA 3934

Query: 5578 GKQTELAVNSGLLTACAFSVDPATTYLEAVKHGAKPVSNCIKMLSNGAGNGQAITTSVDA 5757
            GKQTE   NS LLT C+F+VDPA  YL+AVK GAKPV NC+KML+NG+G+GQAIT ++D+
```

Fig 3 (cont)


```
Sbjct: 3935 GKQTEFVSNSHLLTHCSFAVDPAAAYLDAVKQGAKPVGNCVKMLTNGSGSGQAITCTIDS 3994

Query: 5758 NTNQDSYGGASICLYCRAHVPHPSMDGYCKFKGKCVQVPIGCLDPIRFCLENNVCNVCGC 5937
            NT QD+YGGAS+C+YCRAHV HP+MDG+C++KGK VQVPIG  DPIRFCLEN VC VCGC
Sbjct: 3995 NTTQDTYGGASVCIYCRAHVAHPTMDGFCQYKGKWVQVPIGTNDPIRFCLENTVCKVCGC 4054

Query: 5938 WLGHGCACDRTTIQSVD 5988
            WL HGC CDRT IQS D
Sbjct: 4055 WLNHGCTCDRTAIQSFD 4071
```

4. Sequence D

5325 nucleotides; Replicase

TAGCTTGATTCGTCGAGCAAGGGGTTCTAGTGCAGCTCGACTAGAACCCTGTAATGGCACGGACATCGATAAGTG
TGTTCGTGCTTTTGACATTTATAATAAAAATGTTTCATTCTTGGGTAAGTGTTTGAAGATGAACTGTGTTCGTTT
TAAAAATGCTGATCTTAAGGATGGTTATTTTGTTATAAAGAGGTGTACTAAGTCGGTTATGGAACACGAGCAATC
CATGTATAACCTACTTAACTTTTCTGGTGCTTTGGCTGAGCATGATTTCTTTACTTGGAAAGATGGCAGAGTCAT
TTATGGTAATGTTAGTAGACATAATCTTACTAAATATACTATGATGGACTTGGTTTATGCTATGCGTAACTTTGA
TGAACAAAATTGTGATGTTCTAAAAGAAGTATTAGTTTTAACTGGTTGTTGTGACAATTCTTATTTTGATAGTAA
GGGTTGGTATGACCCAGTTGAAAATGAAGATATACATAGAGTTTATGCATCTCTTGGCAAAATTGTAGCTAGAGC
TATGCTTAAATGCGTTGCTCTATGTGATGCGATGGTTGCTAAAGGTGTTGTTGGTGTTTTAACATTAGATAACCA
AGATCTTAATGGTAACTTTTATGATTTTGGTGATTTTGTTGTTAGCTTACCTAATATGGGTGTTCCCTGTTGTAC
ATCATATTATTCTTATATGATGCCTATTATGGGTTTAACTAATTGTTTAGCTAGTGAGTGTTTTGTCAAGAGTGA
TATTTTTGGTAGTGATTTTAAAACTTTTGATTTGCTTAAGTATGATTTCACTGAACATAAAGAAAATTTATTCAA
TAAGTACTTTAAGCATTGGAGTTTTGATTATCATCCTAATTGTAGTGACTGTTATGATGATATGTGTGTTATACA
TTGTGCTAATTTTAATACACTATTTGCCACAACTATACCAGGTACTGCTTTTGGTCCACTATGTCGTAAAGTTTT
TATAGATGGTGTTCCACTTGTTACAACTGCTGGTTATCATTTTAAGCAATTAGGTTTGGTTTGGAATAAAGATGT
TAACACACACTCAGTTAGGTTGACAATCACTGAACTTTTGCAATTTGTTACTGACCCTTCCTTGATAATAGCTTC
TTCTCCAGCACTCGTTGATCAACGCACTATTTGTTTTTCTGTTGCAGCATTGAGTACTGGTTTGACAAATCAAGT
TGTTAAGCCAGGTCATTTTAATGAAGAGTTTTATAACTTTCTTCGTTTAAGAGGTTTCTTTGATGAAGGTTCTGA
ACTTACATTAAAACATTTCTTCTTCGCACAGAATGGTGATGCTGCTGTTAAAGATTTTGACTTTTACCGTTATAA
TAAGCCTACCATTTTAGATATTTGTCAAGCTAGAGTTACATATAAGATAGTCTCTCGTTATTTTGACATTTATGA
AGGTGGCTGTATTAAGGCATGTGAAGTTGTTGTAACAAATCTTAATAAGAGTGCTGGTTGGCCATTAAATAAGTT
TGGTAAAGCTAGTTTGTATTACGAATCTATATCTTATGAAGAACAGGATGCTTTGTTTGCTTTGACAAAGCGTAA
TGTCCTCCCTACTATGACACAGCTGAATCTTAAGTATGCTATTAGTGGTAAAGAACGTGCTAGAACTGTTGGTGG
TGTTTCTCTGTTGTCCACAATGACCACAAGACAATACCATCAAAAACATCTTAAATCCATTGTTAATACACGCAA
TGCCACTGTTGTTATTGGTACTACCAAATTTTATGGTGGTTGGAATAATATGTTGCGTACTTTAATTGATGGTGT
TGAAAACCCTATGCTCATGGGTTGGGATTATCCCAAATGTGATAGAGCTTTGCCTAACATGATACGTATGATTTC
AGCCATGGTGTTGGGTTCTAAGCATGTTAATTGTTGTACTGTAACAGATAGGTTTTATAGGCTTGGTAACGAGTT
GGCACAAGTTTTAACAGAAGTTGTTTATTCTAATGGTGGTTTTTATTTTAAGCCAGGTGGTACGACTTCTGGTGA
CGCTAGTACAGCTTATGCTAATTCTATTTTTAACATTTTTCAAGCCGTGAGTTCTAACATTAACAGGTTGCTTAG
TGTCCCATCAGATTCATGTAATAATGTTAATGTTAGGGATCTACAACGACGTCTGTATGATAATTGCTATAGGTT
AACTAGTGTTGAAGAGTCATTCATTGATGATTATTATGGTTATCTTAGGAAACATTTTTCAATGATGATTCTCTC
TGATGACGGTGTTGTCTGTTATAACAAGGATTATGCTGAGTTAGGTTATATAGCAGACATTAGTGCTTTTAAAGC
CACTTTGTATTACCAGAATAATGTCTTTATGAGTACTTCTAAATGTTGGGTTGAAGAAGATTTAACTAAGGGACC
ACATGAGTTTTGTTCCCAGCATACTATGCAAATAGTTGATAAAGATGGTACCTATTATTTGCCTTACCCAGATCC
TAGTAGGATCTTGTCAGCTGGTGTTTTTGTTGATGATGTTGTTAAGACAGATGCTGTTGTTTTGTtAGAACGTTA
TGTGTCTTTAGCTATTGATGCATACCCTCTTTCaAAACACCCTAATTCTGAATATCGTAAGGTTTTTTACGTATT
ACTTGATTGGGTTAAGCATCTTAACAAAAATTTGAATGAGGGTGTTCTTGAATCTTTTTCTGTTACACTTCTTGA
TAATCAAGAAGATAAGTTTTGGTGTGAAGATTTTTATGCTAGTATGTATGAAAATTCTACAATATTGCAAGCTGC
TGGCTTATGTGTTGTTTGTGGTTCACAAACTGTTCTTCGTTGTGGTGATTGTCTGCGTAAGCCTATGTTGTGCAC
TAAATGTGCaTATGATCATGTATTTGGTACCGACCACAAGTTTATTTTGGCTATAACACCGTATGTATGTAATGC
ATCAGGTTGTGGTGTTAGTGATGTTAAAAAATTGTATCTTGGTGGTTTGAATTACTATTGTACAAATCATaAACC
ACAGTTGTCTTTTcCATTATGTTCTGCTGGTAATATATTTGGTTTATATAAAAATTCAGCAACTGGTTCCTTAGA
TGTTGAAGTTTTTAATAGGCTTGCAACGTCTGATTGGACTGATGTTAGGGACTATAAACTTGCTAATGATGTTAA
AGATACACTTAGACTCTTTGCGGCTGAAACTATTAAAGCTAAAGAAGAGAGTGTTAAGTCTTCTTATGCTTTTGC
AACTCTTAAAGAGGTTGTTGGACCTAAAGAATTGCTTCTTAGTTGGGAAAGTGGTAAAGTTAAACCACCTTTGAA
TCGTAATTCTGTTTTCACCTGTTTTCAAATAAGTAAGGACTCAAAATTCCAAATAGGTGAGTTCATCTTTGAAAA
GGTTGAATATGGTTCTGATACTGTTACGTATAAGTCTACTGTAACCACTAAGTTAGTTCCTGGTATGATTTTTGT
CTTAACATCTCACAATGTTCAACCTTTACGTGCACCAACTATTGCAAACCAAGAGAAGTATTCTAGCATTTATAA
ATTGCACCCTGCTTTTAATGTCAGTGATGCATATGCTAATTTGGTTCCATATTACCAACTTATTGGTAAACAAAA
GATAACTACAATACAGGGTCCTCCTGGTAGTGGTAAGTCACATTGTTCCATTGGACTTGGATTGTACTATCCAGG
TGCGCGTATTGTTTTTGTTGCTTGTGCCCATGCTGCTGTTGATTCCTTATGTGCAAAAGCTATGACTGTTTATAG
CATTGATAAGTGTACTAGGATTATACCTGCAAGAGCTCGGGTTGAGTGTTATAGTGGCTTTAAACCAAATAACAC
TAGTGCACAATACATATTTAGCACTGTTAACGCATTACCTGAGTGTAATGCTGATATTGTTGTTGTAGATGAAGT
TTCAATGTGTACAAATTATGACCTTTCTGTTATTAATCAGCGTTTATCATATAAACATATTGTTTATGTTGGTGA
TCCACAACAACTTCCTGCACCTAGAGTAATGATTACTAAAGGTGTTATGGAGCCTGTTGATTATAACGTTGTTAC
TCAACGTATGTGTGCTATAGGCCCTGATGTTTTTCTTCATAAATGTTATAGATGTCCTGCTGAAATAGTTAATAC

Fig 3 (Cont) 15/25

AGTTTCTGAACTTGTTTATGAGAACAAGTTTGTCCCTGTTAAACCTGCTAGTAAACAGTGTTTTAAAATCTTTTT
TAAGGGTAATGTACAGGTTGACAATGGCTCTAGTATTAACAGAAAGCAGCTTGAAATAGTTAAGCTGTTTTTAGT
TAAAAATCCAAGTTGGAGTAAGGCTGTGTTTATTTCTCCTTATAATAGTCAGAATTATGTTGCTAGTAGATTTTT
AGGACTTCAAATTCAAACTGTTGATTCTTCTCAAGGTAGTGAGTATGATTATGTAATCTATGCACAAACTTCTGA
CACTGCACATGCTTGCAATGTAAACCGTTTTAATGTTGCTATAACACGTGCTAAGAAGGGTATATTTTGTGTAAT
GTGTGATAAAACTTTGTTTGATTCACTTAAGTTTTTTGAGATTAAACATGCAGATTTACACTCTAGCCAGGTTTC
TGGCTTGTTTAAAAATTGTACACGCACTCCTCTTAATTTACCACCAACTCATGCACACACTTTCTTGTCGTTGTC
AGATCAGTTTAAGACTACAGGTGATTTAGCTGTTCAAATAGGTTCAAATAATGTTTGTACTTATGAACATGTTAT
ATCATTTATGGGTTTTAGGTTTGATATTAGTATTCCTGGTAGTCATAGTTTGTTTTGTACACGTGACTTTGCTAT
TCGTAATGTGCGTGGTTGGTTGGGTATGGATGTTGAAAGTGCTCATGTTTGTGGCGATAACATAGGTACTAATGT
TCCTTTACAGGTTGGTTTTTCAAATGGTGTTAATTTTGTTGTGCAAACTGAAGGTTGTGTGTCTACCAATTTTGC
TGATGTTATTAAACCTGTTTGTGCAAAATCTCCACCAGGTGAACAATTTAGACACCTTGTTCCTTTTTTACGTAA
AGGACAACCTTGGTTAATTGTTCGTAGACGCATTGTGCAAATGATATCTGATTATTTGTCCAATTTGTCTGACAT
TCTTGTCTTTGTTTTGTGGGCAGGTAGTTTGGAATTAACTACAATGCGTTACTTTGTAAAAATAGGGCCAATTAA
ATATTGTTATTGTGGTAATTCTGCCACTTGTTATAATTCAGTTAGTAATGAATATTGTTGTTTTAAACATGCATT
GGGTTGTGATTATGTTTACAATCCGTATGCTTTTGATATACAACAGTGGGGTTATGTTGGTTCCTTGAGCCAGAA

Hypothesized ORFs

>~out: -1 to 5320: Frame 2 1774 aa

SLIRRARGSSAARLEPCNGTDIDKCVRAFDIYNKNVSFLGKCLKMNCVRFKNADLKDGYFVIKRCTKSVMEHEQS
MYNLLNFSGALAEHDFFTWKDGRVIYGNVSRHNLTKYTMMDLVYAMRNFDEQNCDVLKEVLVLTGCCDNSYFDSK
GWYDPVENEDIHRVYASLGKIVARAMLKCVALCDAMVAKGVVGVLTLDNQDLNGNFYDFGDFVVSLPNMGVPCCT
SYYSYMMPIMGLTNCLASECFVKSDIFGSDFKTFDLLKYDFTEHKENLFNKYFKHWSFDYHPNCSDCYDDMCVIK
CANFNTLFATTIPGTAFGPLCRKVFIDGVPLVTTAGYHFKQLGLVWNKDVNTHSVRLTITELLQFVTDPSLIIAS
SPALVDQRTICFSVAALSTGLTNQVVKPGHFNEEFYNFLRLRGFFDEGSELTLKHFFFAQNGDAAVKDFDFYRYN
KPTILDICQARVTYKIVSRYFDIYEGGCIKACEVVVTNLNKSAGWPLNKFGKASLYYESISYEEQDALFALTKRN
VLPTMTQLNLKYAISGKERARTVGGVSLLSTMTTRQYHQKHLKSIVNTRNATVVIGTTKFYGGWNNMLRTLIDGV
ENPMLMGWDYPKCDRALPNMIRMISAMVLGSKHVNCCTVTDRFYRLGNELAQVLTEVVYSNGGFYFKPGGTTSGD
ASTAYANSIFNIFQAVSSNINRLLSVPSDSCNNVNVRDLQRRLYDNCYRLTSVEESFIDDYYGYLRKHFSMMILS
DDGVVCYNKDYAELGYIADISAFKATLYYQNNVFMSTSKCWVEEDLTKGPHEFCSQHTMQIVDKDGTYYLPYPDE
SRILSAGVFVDDVVKTDAVVLLERYVSLAIDAYPLSKHPNSEYRKVFYVLLDWVKHLNKNLNEGVLESFSVTLLD
NQEDKFWCEDFYASMYENSTILQAAGLCVVCGSQTVLRCGDCLRKPMLCTKCAYDHVFGTDHKFILAITPYVCNA
SGCGVSDVKKLYLGGLNYYCTNHKPQLSFPLCSAGNIFGLYKNSATGSLDVEVFNRLATSDWTDVRDYKLANDVK
DTLRLFAAETIKAKEESVKSSYAFATLKEVVGPKELLLSWESGKVKPPLNRNSVFTCFQISKDSKFQIGEFIFEK
VEYGSDTVTYKSTVTTKLVPGMIFVLTSHNVQPLRAPTIANQEKYSSIYKLHPAFNVSDAYANLVPYYQLIGKQK
ITTIQGPPGSGKSHCSIGLGLYYPGARIVFVACAHAAVDSLCAKAMTVYSIDKCTRIIPARARVECYSGFKPNNT
SAQYIFSTVNALPECNADIVVVDEVSMCTNYDLSVINQRLSYKHIVYVGDPQQLPAPRVMITKGVMEPVDYNVVT
QRMCAIGPDVFLHKCYRCPAEIVNTVSELVYENKFVPVKPASKQCFKIFFKGNVQVDNGSSINRKQLEIVKLFLV
KNPSWSKAVFISPYNSQNYVASRFLGLQIQTVDSSQGSEYDYVIYAQTSDTAHACNVNRFNVAITRAKKGIFCVM
CDKTLFDSLKFFEIKHADLHSSQVCGLFKNCTRTPLNLPPTHAHTFLSLSDQFKTTGDLAVQIGSNNVCTYEHVI
SFMGFRFDISIPGSHSLFCTRDFAIRNVRGWLGMDVESAHVCGDNIGTNVPLQVGFSNGVNFVVQTEGCVSTNFG
DVIKPVCAKSPPGEQFRHLVPFLRKGQPWLIVRRRIVQMISDYLSNLSDILVFVLWAGSLELTTMRYFVKIGPIK
YCYCGNSATCYNSVSNEYCCFKHALGCDYVYNPYAFDIQQWGYVGSLSQ

>~out: 189 to 341: Frame 3 51 aa

RGVLSRLWNTSNPCITYLTFLVLWLSMISLLGKMAESFMVMLVDIILLNIL

>~out: 726 to 977: Frame 3 84 aa

LVSVLSRVIFLVVILKLLICLSMISLNIKKIYSISTLSIGVLIIILIVVTVMMICVLYIVLILIHYLPQLYQVLLLV
HYVVKFL

>~out: 2661 to 2903: Frame 3 81 aa

MRVFLNLFLLHFLIIKKISFGVKIFMLVCMKILQYCKLLAYVLFVVHKLFFVVVIVCVSLCCALNVHMIMYI
VPTTSLFWL

>~out: 3075 to 3296: Frame 3 74 aa

MLKFLIGLQRLIGLMLGTINLLMMLKIHLDSLRLKLLKLKKRVLSLLMLLQLLKRLLDLKNCFLVGKVVK
LNHL

>~out: 3741 to 3890: Frame 3 50 aa

LFIALISVLGLYLQELGLSVIVALNQITLVHNTYLALLTHYLSVMLILLL

>~out: 4500 to 4676: Frame 3 59 aa

CVIKLCLIHLSFLRLNMQIYTLARFVACLKIVHALLLIYHQLMHTLSCRCQISLRLQVI

>~out: 4692 to 4862: Frame 3 57 aa

VQIMFVLMNMLYHLWVLGLILVFLVVIVCFVHVTLLFVMCVVGWVWMLKVLMFVAIT

>~out: 4866 to 5039: Frame 3 58 aa

VLMFLYRLVFQMVLILLCKLKVVCLPILVMLLNLFVQNLHQVNNLDTLFLFYVKDNLG

>~out: 5166 to 5315: Frame 3 50 aa

GQLNIVIVVILPLVIIQLVMNIVVLNMHWVVIMFTIRMLLIYNSGVMLVP

Fig 3 (cont)

Alignment

>gi|12175747|ref|NP_073549.1| replicase polyprotein 1ab [Human coronavirus 229E]
gi|30179827|sp|Q05002|R1AB_CVH22 Replicase polyprotein 1ab (pp1ab) (ORF1ab polyprotein) [Includes: Replicase polyprotein 1a (pp1a) (ORF1a)] [Contains: p9; p87; p195 (Papain-like proteinases 1/2) (PL1-PRO/PL2-PRO); Peptide HD2; 3C-like proteinase (3CL-PRO) (3CLp) (M-PRO) (p34); Unknown protein 1; p5; p23; p12; Growth factor-like peptide (GFL) (p16); RNA-directed RNA polymerase (RdRp) (Pol) (p100); Helicase (Hel) (p66) (p66-HEL); Unknown protein 2; p41; Unknown protein 3]
gi|12082740|gb|AAG48591.1| replicase polyprotein 1ab [Human coronavirus 229E]
Length = 6758

Score = 3137 bits (8134), Expect = 0.0
Identities = 1465/1773 (82%), Positives = 1633/1773 (92%)
Frame = +2

```
Query: 2    SLIRRARGSSAARLEPCNGTDIDKCVRAFDIYNKNVSFLGKCLKMNCVRFKNADLKDGYF 181
            S + R RGSSAARLEPCNGTDID CVRAFD+YNK+ SF+GK LK NCVRFKN D  D ++
Sbjct: 4073 SYLNRVRGSSAARLEPCNGTDIDYCVRAFDVYNKDASFIGKNLKSNCVRFKNVDKDDAFY 4132

Query: 182  VIKRCTKSVMEHEQSMYNLLNFSGALAEHDFFTWKDGRVIYGNVSRHNLTKYTMMDLVYA 361
            ++KRC KSVM+HEQSMYNLL     A+A+HDFFTW +GR IYGNVSR +LTKYTMMDL +A
Sbjct: 4133 IVKRCIKSVMDHEQSMYNLLKGCNAVAKHDFFTWHEGRTIYGNVSRQDLTKYTMMDLCFA 4192

Query: 362  MRNFDEQNCDVLKEVLVLTGCCDNSYFDSKGWYDPVENEDIHRVYASLGKIVARAMLKCV 541
            +RNFDE++C+V KE+LVLTGCC    YF+ K W+DP+ENEDIHRVYA+LGK+VA AMLKCV
Sbjct: 4193 LRNFDEKDCEVFKEILVLTGCCSTDYFEMKNWFDPIENEDIHRVYAALGKVVANAMLKCV 4252

Query: 542  ALCDAMVAKGVVGVLTLDNQDLNGNFYDFGDFVVSLPNMGVPCCTSYYSYMMPIMGLTNC 721
            A CD MV KGVVGVLTLDNQDLNGNFYDFGDFV+   P MG+P CTSYYSYMMP+MG+TNC
Sbjct: 4253 AFCDEMVLKGVVGVLTLDNQDLNGNFYDFGDFVLCPPGMGIPYCTSYYSYMMPVMGMTNC 4312

Query: 722  LASECFVKSDIFGSDFKTFDLLKYDFTEHKENLFNKYFKHWSFDYHPNCSDCYDDMCVIH 901
            LASECF+KSDIFG DFKTFDLLKYDFTEHKE LFNKYFK+W  DYHP+C DC+D+MC++H
Sbjct: 4313 LASECFMKSDIFGQDFKTFDLLKYDFTEHKEVLFNKYFKYWGQDYHPDCVDCHDEMCILH 4372

Query: 902  CANFNTLFATTIPGTAFGPLCRKVFIDGVPLVTTAGYHFKQLGLVWNKDVNTHSVRLTIT 1081
            C+NFNTLFATTIP TAFGPLCRKVFIDGVP+V TAGYHFKQLGLVWNKDVNTHS RLTIT
Sbjct: 4373 CSNFNTLFATTIPNTAFGPLCRKVFIDGVPVVATAGYHFKQLGLVWNKDVNTHSTRLTIT 4432

Query: 1082 ELLQFVTDPSLIIASSPALVDQRTICFSVAALSTGLTNQVVKPGHFNEEFYNFLRLRGFF 1261
            ELLQFVTDP+LI+ASSPALVD+RT+CFSVAALSTGLT+Q VKPGHFN+EFY+FLR +GFF
Sbjct: 4433 ELLQFVTDPTLIVASSPALVDKRTVCFSVAALSTGLTSQTVKPGHFNKEFYDFLRSQGFF 4492

Query: 1262 DEGSELTLKHFFFAQNGDAAVKDFDFYRYNKPTILDICQARVTYKIVSRYFDIYEGGCIK 1441
            DEGSELTLKHFFF Q GDAA+KDFD+YRYN+PT+LDI QARV Y++  +RYFD YEGGCI
Sbjct: 4493 DEGSELTLKHFFFTQKGDAAIKDFDYYRYNRPTMLDIGQARVAYQVAARYFDCYEGGCIT 4552

Query: 1442 ACEVVVTNLNKSAGWPLNKFGKASLYYESISYEEQDALFALTKRNVLPTMTQLNLKYAIS 1621
            + EVVVTNLNKSAGWPLNKFGKA LYYESISYEEQDA+F+LTKRN+LPTMTQLNLKYAIS
Sbjct: 4553 SREVVVTNLNKSAGWPLNKFGKAGLYYESISYEEQDAIFSLTKRNILPTMTQLNLKYAIS 4612

Query: 1622 GKERARTVGGVSLLSTMTTRQYHQKHLKSIVNTRNATVVIGTTKFYGGWNNMLRTLIDGV 1801
            GKERARTVGGVSLL+TMTTRQ+HQK LKSIV TRNATVVIGTTKFYGGW+NML+ L+  V
Sbjct: 4613 GKERARTVGGVSLLATMTTRQFHQKCLKSIVATRNATVVIGTTKFYGGWDNMLKNLMADV 4672

Query: 1802 ENPMLMGWDYPKCDRALPNMIRMISAMVLGSKHVNCCTVTDRFYRLGNELAQVLTEVVYS 1981
            ++P LMGWDYPKCDRA+P+MIRM+SAM+LGSKHV CCT +D+FYRL NELAQVLTEVVYS
Sbjct: 4673 DDPKLMGWDYPKCDRAMPSMIRMLSAMILGSKHVTCCTASDKFYRLSNELAQVLTEVVYS 4732

Query: 1982 NGGFYFKPGGTTSGDASTAYANSIFNIFQAVSSNINRLLSVPSDSCNNVNVRDLQRRLYD 2161
            NGGFYFKPGGTTSGDA+TAYANS+FNIFQAVSSNIN +LSV S +CNN NV+ LQR+LYD
Sbjct: 4733 NGGFYFKPGGTTSGDATTAYANSVFNIFQAVSSNINCVLSVNSSNCNNFNVKKLQRQLYD 4792

Query: 2162 NCYRLTSVEESFIDDYYGYLRKHFSMMILSDDGVVCYNKDYAELGYIADISAFKATLYYQ 2341
            NCYR ++V+ESF+DD+YGYL+KHFSMMILSDD VVCYNK YA LGYIADISAFKATLYYQ
Sbjct: 4793 NCYRNSNVDESFVDDFYGYLQKHFSMMILSDDSVVCYNKTYAGLGYIADISAFKATLYYQ 4852

Query: 2342 NNVFMSTSKCWVEEDLTKGPHEFCSQHTMQIVDKDGTYYLPYPDPSRILSAGVFVDDVVK 2521
            N VFMST+KCW EEDL+ GPHEFCSQHTMQIVD++G YYLPYPDPSRI+SAGVFVDD+ K
Sbjct: 4853 NGVFMSTAKCWTEEDLSIGPHEFCSQHTMQIVDENGKYYLPYPDPSRIISAGVFVDDITK 4912
```

```
Query: 2522 TDAVVLLERYVSLAIDAYPLSKHPNSEYRKVFYVLLDWVKHLNKNLNEGVLESFSVTLLD 2701
             TDAV+LLERYVSLAIDAYPLSKHP  EYRKVFY LLDWVKHLNK LNEGVLESFSVTLLD
Sbjct: 4913 TDAVILLERYVSLAIDAYPLSKHPKPEYRKVFYALLDWVKHLNKTLNEGVLESFSVTLLD 4972

Query: 2702 NQEDKFWCEDFYASMYENSTILQAAGLCVVCGSQTVLRCGDCLRKPMLCTKCAYDHVFGT 2881
               E KFW E FYASMYE ST+LQAAGLCVVCGSQTVLRCGDCLR+PMLCTKCAYDHVFGT
Sbjct: 4973 EHESKFWDESFYASMYEKSTVLQAAGLCVVCGSQTVLRCGDCLRRPMLCTKCAYDHVFGT 5032

Query: 2882 DHKFILAITPYVCNASGCGVSDVKKLYLGGLNYYCTNHKPQLSFPLCSAGNIFGLYKNSA 3061
             DHKFILAITPYVCN SGC V+DV KLYLGGLNYYC +HKP LSFPLCSAGN+FGLYK+SA
Sbjct: 5033 DHKFILAITPYVCNTSGCNVNDVTKLYLGGLNYYCVDHKPHLSFPLCSAGNVFGLYKSSA 5092

Query: 3062 TGSLDVEVFNRLATSDWTDVRDYKLANDVKDTLRLFAAETIKAKEESVKSSYAFATLKEV 3241
              GS+D++VFN+L+TSDW+D+RDYKLAND K++LRLFAAET+KAKEESVKSSYA+ATLKE+
Sbjct: 5093 LGSMDIDVFNKLSTSDWSDIRDYKLANDAKESLRLFAAETVKAKEESVKSSYAYATLKEI 5152

Query: 3242 VGPKELLLSWESGKVKPPLNRNSVFTCFQISKDSKFQIGEFIFEKVEYGSDTVTYKSTVT 3421
             VGPKELLL WESGK KPPLNRNSVFTCFQI+KDSKFQ+GEF+FEKV+YGSDTVTYKST T
Sbjct: 5153 VGPKELLLLWESGKAKPPLNRNSVFTCFQITKDSKFQVGEFVFEKVDYGSDTVTYKSTAT 5212

Query: 3422 TKLVPGMIFVLTSHNVQPLRAPTIANQEKYSSIYKLHPAFNVSDAYANLVPYYQLIGKQK 3601
             TKLVPGM+F+LTSHNV PLRAPT+ANQEKYS+IYKLHP+FNVSDAYANLVPYYQLIGKQ+
Sbjct: 5213 TKLVPGMLFILTSHNVAPLRAPTMANQEKYSTIYKLHPSFNVSDAYANLVPYYQLIGKQR 5272

Query: 3602 ITTIQGPPGSGKSHCSIGLGLYYPGARIVFVACAHAAVDSLCAKAMTVYSIDKCTRIIPA 3781
             ITTIQGPPGSGKSHCSIG+G+YYPGARIVF AC+HAAVDSLCAKA+T YS+DKCTRIIPA
Sbjct: 5273 ITTIQGPPGSGKSHCSIGIGVYYPGARIVFTACSHAAVDSLCAKAVTAYSVDKCTRIIPA 5332

Query: 3782 RARVECYSGFKPNNTSAQYIFSTVNALPECNADIVVVDEVSMCTNYDLSVINQRLSYKHI 3961
             RARVECYSGFKPNN SAQY+FSTVNALPE NADIVVVDEVSMCTNYDLSVINQR+SYKHI
Sbjct: 5333 RARVECYSGFKPNNNSAQYVFSTVNALPEVNADIVVVDEVSMCTNYDLSVINQRISYKHI 5392

Query: 3962 VYVGDPQQLPAPRVMITKGVMEPVDYNVVTQRMCAIGPDVFLHKCYRCPAEIVNTVSELV 4141
             VYVGDPQQLPAPRV+I+KGVMEP+DYNVVTQRMCAIGPDVFLHKCYRCPAEIVNTVSELV
Sbjct: 5393 VYVGDPQQLPAPRVLISKGVMEPIDYNVVTQRMCAIGPDVFLHKCYRCPAEIVNTVSELV 5452

Query: 4142 YENKFVPVKPASKQCFKIFFKGNVQVDNGSSINRKQLEIVKLFLVKNPSWSKAVFISPYN 4321
             YENKFVPVK ASKQCFKIF +G+VQVDNGSSINR+QL++VK F+ KN +WSKAVFISPYN
Sbjct: 5453 YENKFVPVKEASKQCFKIFERGSVQVDNGSSINRRQLDVVKRFIHKNSTWSKAVFISPYN 5512

Query: 4322 SQNYVASRFLGLQIQTVDSSQGSEYDYVIYAQTSDTAHACNVNRFNVAITRAKKGIFCVM 4501
             SQNYVA+R LGLQ QTVDS+QGSEYDYVI+AQTSDTAHACN NRFNVAITRAKKGIFC+M
Sbjct: 5513 SQNYVAARLLGLQTQTVDSAQGSEYDYVIFAQTSDTAHACNANRFNVAITRAKKGIFCIM 5572

Query: 4502 CDKTLFDSLKFFEIKHADLHSSQVCGLFKNCTRTPLNLPPTHAHTFLSLSDQFKTTGDLA 4681
              D+TLFD+LKFFEI    DL S    CGLFK+C R P++LPP+HA T+LSLSD+FKT+GDLA
Sbjct: 5573 SDRTLFDALKFFEITMTDLQSESSCGLFKDCARNPIDLPPSHATTYLSLSDRFKTSGDLA 5632

Query: 4682 VQIGSNNVCTYEHVISFMGFRFDISIPGSHSLFCTRDFAIRNVRGWLGMDVESAHVCGDN 4861
             VQIG+NNVCTYEHVIS+MGFRFD+S+PGSHSLFCTRDFA+R+VRGWLGMDVE AHV GDN
Sbjct: 5633 VQIGNNNVCTYEHVISYMGFRFDVSMPGSHSLFCTRDFAMRHVRGWLGMDVEGAHVTGDN 5692

Query: 4862 IGTNVPLQVGFSNGVNFVVQTEGCVSTNFGDVIKPVCAKSPPGEQFRHLVPFLRKGQPWL 5041
             +GTNVPLQVGFSNGV+FV Q EGCV TN G V+KPV A++PPGEQF H+VP LRKGQPW
Sbjct: 5693 VGTNVPLQVGFSNGVDFVAQPEGCVLTNTGSVVKPVRARAPPGEQFTHIVPLLRKGQPWS 5752

Query: 5042 IVRRRIVQMISDYLSNLSDILVFVLWAGSLELTTMRYFVKIGPIKYCYCGNSATCYNSVS 5221
             ++R+RIVQMI+D+L+  SD+LVFVLWAG LELTTMRYFVKIG +K+C CG  ATCYNSVS
Sbjct: 5753 VLRKRIVQMIADFLAGSSDVLVFVLWAGGLELTTMRYFVKIGAVKHCQCGTVATCYNSVS 5812

Query: 5222 NEYCCFKHALGCDYVYNPYAFDIQQWGYVGSLS 5320
             N+YCCFKHALGCDYVYNPY  DIQQWGYVGSLS
Sbjct: 5813 NDYCCFKHALGCDYVYNPYVIDIQQWGYVGSLS 5845
```

5. Sequence E

6143 nucleotides; 3' end of Replicase and 5' end of Spike

TCTGGAATTGTAATGTTgATATGTATCCAGAATTTTCAATTGTGTGTCGCTTTGACACACGTACTCGTTCTGTTT
TTAATTTAGAAGGTGTTAATGGTGGTTCTCTTTATGTTAACAAACATGCGTTTCATACACCAGCATATGATAAAC
GTGCTTTTGTTAAATTAAAACCTATGCCCTTTTTTTACTTTGATGACAGTGATTGTGATGTTGTGCAAGAACAAG
TTAATTATGTACCCCTTCGCGCTAGTAGTTGTGTTACCCGTTGTAATATAGGTGGTGCTGTTTGTTCAAAACATG
CAAATTTGTATCAAAAATATGTTGAGGCATATAATACATTTACACAGGCTGGTTTTAACATTTGGGTACCACATA
GTTTTGATGTTTATAATTTGTGGCAAATTTTTATTGAAACTAATTTACAAAGTCTTGAAAATATAGCATTTAATG
TTGTAAAAAAaGGGTGTTTTACTGGTGTTGATGGTGAGTTACCTGTTGCAGTTGTTAACGACAAAGTTTTTGTTC
GCTATGGCGATGTTGACAACTTGGTTTTTACAAATAAAACAACATTGCCTACTAATGTTGCTTTTGAATTGTTTG

CAAAACGAAAAATGGGTTTAACACCACCATTGTCTATTCTCAAAAATCTTGGTGTTGTTGCTACATATAAATTTG
TTTTATGGGATTATGAAGCTGAAAGACCTTTTACCTCATATACTAAGAGTGTATGTAAATACACTGATTTTAATG
AGGATGTTTGTGTTTGTTTTGACAATAGTATTCAGGGTTCGTATGAGCGTTTTACGCTTACTACGAACGCTGTTT
TATTTTCTACTGTTGTCATTAAAAATTTAACACCTATAAAGTTGAATTTTGGTATGTTGAATGGTATGCCAGTTT
CTTCTATTAAGAGTGATAAAGGTGTTGAAAAATTAGTTAATTGGTACAYATATGTTCGTAAAAATGGTCAATTTC
AAGATCATTATGATGGTTTTTACACTCAAGGTAGGAATTTATCAGACTTTACACCAAGAAGTGATATGGAGTATG
ATTTTCTTAACATGGATATGGGTGTTTTTATTAATAAATATGGTCTTGAGGATTTTAATTTTGAACATGTTGTAT
ATGGTGATGTTTCAAAAACTACATTAGGAGGTCTTCATTTGTTGATATCACAGTTTAGGCTTAGTAAAATGGGTG
TTTTGAAAGCTGATGATTTTGTCACTGCTTCTGACACAACTTTGAGGTGCTGTACTGTTACTTATCTTAATGAAC
TTAGTTCAAAAGTTGTTTGTACTTATATGGATTTGTTGTTGGACGACTTTGTTACTATACTAAAGAGTTTAGATC
TTGGTGTAATATCTAAAGTTCATGAAGTTATTATAGATAATAAACCTTATAGGTGGATGTTGTGGTGTAAAGATA
ACCACTTGTCCACTTTTTATCCACAGTTGCAGTCTGCTGAATGGAAGTGTGGTTATGCTATGCCACAAATTTATA
AGCTTCAACGWATGTGTTTGGAACCTTGTAATTTATATAATTATGGTGCTGGTATTAAGTTGCCTAGTGGTATAA
TGTTAAATGTTGTTAAATACACTCAGCTTTGTCAATACCTAAATAGCACTACAATGTGCGTACCTCATAATATGC
GTGTTTTGCACTATGGTGCTGGTTCTGACAAAGGTGTGGCACCTGGTACAACTGTTTTAAAACGTTGGCTACCAC
CTGATGCAATAATCATTGATAATGATATCAATGATTATGTTAGTGATGCAGATTTTAGCATTACAGGTGATTGTG
CTACTGTTTACCTTGAAGATAAGTTTGACTTACTTATTTCTGATATGTATGATGGTAGAATTAAATTTTGTGATG
GTGAAAACGTCTCTAAAGATGGTTTTTTTACTTATCTTAATGGTGTTATTAGAGAAAAATTAGCTATTGGTGGTA
GTGTTGCCATTAAGATTACAGAATATAGTTGGAATAAGTATCTTTATGAATTAATACAAAGATTTGCTTTTTGGA
CTTTGTTCTGCACGTCTGTTAATACATCCTCTTCAGAAGCTTTTCTTATTGGTATTAATTATTTAGGTGACTTTA
TTCAAGGTCCTTTTATAGCTGGTAACACTGTTCATGCTAATTATATATTTTGGCGTAATTCTACTATTATGTCTT
TGTCATACAATTCAGTTTTAGATTTAAGTAAGTTTGAATGTAAACATAAGGCCACTGTTGTTGTTACACTTAAAG
ATAGTGATGTAAATGATATGGTTTTGAGTTTGATTAAGAGTGGTAGGTTGTTGTTACGTAATAGTGGCCGTTTTG
GTGGTTTTAGTAATCATTTAGTCTCAACTAAATGAAACTTTTCTTGATTTTGCTTATTTTGCCCCTGGTTTCTTG
CTTTTCTACATGTAACAGTAATGCTAGTATTTCTATGTTACAATTAGGTGTTCCTGATAACTCTTCAACTATTGT
CACAGGTTTGTTGCCAGTCCATTGGATTTGTGCTAATCAGAGTACATCTAGTTACCCAGCCAACGGCTTTTTCTA
TATTGATGTTGGTAAACACCGTAGTGCCTTTGCACTCCATAGTGGTTATTATGATGCTAACCAGTATTATATTTA
TCTCACTAATAAAATACATTTAAATGCTCCTGTCACTCTGAAGATTTGTAAGTTTGGAAACACTTCTTTTGATTT
TTTAAGTAATGTTTCTACTTCTCATGATTGTATAGTTAATTTGTCATTCACAGAACAGTTAGGTGTGCCTTTGGG
CATAACTATATCGGGTGAAACTGTACGTTTGCATTTATATAATGCAACTCGTACTTTTTATGTGCCGGCCGCTTA
TAAACTTACTAAACTTAGTGTTAAATGTTACTTTAGTGAATCCTGTGTTTTTAGTGTTGTCAATGCCACCATTAC
TGTTAATGTCACCACACTTAATGGCCGTATAGTTAACTACACTGTTTGTGATGATTGTAATGGTTATACTGATAA
CATATTTTCTGTTCAACAGGATGGCCGCATTCCTAATGGTTTCCCTTTTAATAATTGGTTTTTGTTAACTAATGG
TTCCACATTAGTGGACGGGGTCTCTAGACTTTATCAACCACTCCGTTTAACTTGTTTATGGCCTGTACCTGGTCT
TAAATCTTCAACTGGTTTTGTTTATTTTAATGCCACTGGTTCTGATGTTAATTGTAACGGCTATCAACATAATTC
TGTTGCTGATGTTATGCGTTACAATCTTAACCTCAGTGCTAATTCTGTGGACAATCTTAAGAGTGGTGTTATAGT
TTTTAAAACTTTACAGTACGATGTTTTGTTTTATTGTAGTAATTCTTCTTCAGGTGTTCTTGACACCACAATACC
TTTTGGCCCTTCCTCTCAACCTTATTACTGTTTTATAAACAGTACTATCAACACTACTCATGTTAGCACTTTTGT
GGGTATTTTACCACCCACTGTGCGTGAAATTGTTGTTGCTAGAACTGGTCAGTTTTATATTAATGGTTTTAAGTA
TTTCGATTTGGGTTTCATAGAAGCTGTCAATTTTAATGTCACGACTGCTAGTGCCACAGATTTTTGGACGGTTGC
ATTTGCTACTTTTGTTGATGTTTTGGTTAATGTTAGTGCAACTAACATTCAAAACTTACTTTATTGCGATTCTCC
ATTTGAAAAGTTGCAGTGTGAGCACTTGCAGTTTGGATTGCAAGATGGTTTTTATTCTGCAAATTTTCTTGATGA
TAATGTTTTGCCTGAGACTTATGTTGCACTCCCCATTTATTATCAACATACGGACATAAATTTTACTGCAACTGC
ATCTTTTGGTGGTTCTTGTTATGTTTGTAAACCACGCCAGGTTAATATATCTCTTAATGGTAACACTTCAGTGTG
TGTTAGAACATCTCATTTTTCAATTAGGTATATTTATAACCGCGTTAAGAGTGGTTCACCAGGTGACTCTTCATG
GCATATTTATTTAAAGAGTGGCACTTGTCCATTTTCTTTTTCTAAGTTAAATAATTTTCAAAAGTTTAAGACTAT
TTGTTTCTCAACCGTCGAAGTGCCTGGTAGTTGTAATTTTCCACTTGAAGCCACCTGGCATTACACTTCTTATAC
TATTGTTGGTGCTTTGTATGTTACTTGGTCTGAAGGTAATTCCATTACTGGTGTACCTTATCCTGTCTCTGGTAT
TCGTGAGTTTAGTAATTTAGTTTTAAATAATTGTACCAAATATAATATTTATGATTATGTTGGTACTGGAATTAT
ACGTTCTTCAAACCAGTCACTTGCTGGTGGTATTACATATGTTTCTAACTCTGGTAATTTACTTGGTTTTAAAAA
TGTTTCCACTGGTAACATTTTTATTGTGACACCATGTAACCAACCAGATCAAGTAGCTGTTTATCAACAAAGCAT
TATTGGTGCCATGACCGCTGTTAATGAGTCTAGATATGGCTTGCAAAACTTACTACAGTTACCTAACTTTTATTA
TGTTAGTAATGGTGGTAACAATTGCACTACGGCTGTTATGATTTATTCTAATTTTGGTATTTGTGCTGATGGTTC
TTTAATTCCTGTtCGTCCGCGTAATTCTAGTGATAATGGTATTTCAGCCATAATCACTGCTAATTTATCCATTCC
CTCTAACTGGACTACTTCAGTTCAAGTTGAGTACCTCCAAATTACTAGTACTCCAATAGTTGTTGATTGTGCTAC
TTATGTGTGTAATGGTAACCCTCGTTGTAAGAATCTACTTAAGCAGTATACTTCTGCTTGTAAAACTATTGAAGA
TGCCTTACGACTTAGTGCTCATTTGGAAACTAATGATGTTAGTAGTATGCTAACTTTCGATAGCAATGCTTTTAG
TTTGGCTAATGTTACTAGTTTTGGAGATTATAACCTTTCTAGTGTTTTACCTCAGAGAAACATTCATTCAAGCCG
TATAGCAGGACGTAGTGCTTTGGAAGATTTGTTGTTTAGCAAAGTTGTTACATCTGGTTTGGGTACTGTTGATGT
TGACTATAAGTCTTGTACTAAAGGTCTTTCTATTGCTGACCTTGCTTGTGCTCAGTACTACAATGGCATAATGGT
TTTGCCAGGTGTTGCTGATGCTGAACGTATGGCCATGTACACAGGTTCTCTTATAGGTGGCATGGTGCTCGGAGG
TCTTACATCAGCAGCCGCCATACCTTTTTCTTTGGCACTGCAAGCACGACTTAACTATGTTGCTTTACAAACTGA
TGTGCTTCAAGAAAATCAGAAAATTTTGGCTGCATCATTTAATAAGGCTATTAATAATATTGTTGCTTCTTTTAG
TAGCGTTAATGATGCTATTACACATACTGCAGAGGCTATACATACTGTTACTATTGCACTTAATAAGATTCAGGA
TGTTGTTAATCAACAGGGTAGTGCTCTTAACCATCTCACTTCACAATTGAGACATAATTTTCAGGCCATTTCTAA
TTCAATTCATGCTATTTATGACCGGCTTGATTCAATTCAAGCCGATCAACAAGTTGACAGATTAATTACTGGACG
GCTTGCAGCTTTGAATGCATTTGTTTCCCAAGTTTTGAATAAATATACTGAAGTTCGTGGTTCCAGACGCTTAGC
ACAGCAGAAGATTAATGAATGTGTCAAGTCACAATCTAATAGATATGGTTTTTGTGGCAATGGCACTCACATCTT

Fig 3. (Cont.) 19/25

TTCAATCGTCAACTCAGCTCCAGATGGTTTGCTTTTTCTTCATACTGTTTTGCTGCCAACTGATTACAAGAATG
AAAGGCGTGGTCTGGTATCTGTGTTGATGGCATTTATGGCTATGTTCTGCGTCAACCTAACTTGGTTCTTTATT
TGATAATGGTGTCTTTCGTGTAACTTCCAGGGTCATGTTTCAACCTCGTTTACCTGTTTTGTCTGATTTTGTGC
AATATATAATTGTAATGTTACTTTTGTTAACATATCTCGTGTCGAGTTACATACTGTCATACCTGACTACGTTG
TGTTAATAAAACATTACAAGAGTTTGCACAAAACTTACCAAAGTATGTTAAGCCTAATTTTGACTTGACTCCTT
TAATTTAACATATCTTAATTTGAGTTCTGAGTTGAAGCAACTCGAAGCTAAAACTGCTACGAATCAGC

Hypothesised ORFs

>~out: 3 to 2357: Frame 3 785 aa
WNCNVDMYPEFSIVCRFDTRTRSVFNLEGVNGGSLYVNKHAFHTPAYDKRAFVKLKPMPFFYFDDSDCDVVQEQ
NYVPLRASSCVTRCNIGGAVCSKHANLYQKYVEAYNTFTQAGFNIWVPHSFDVYNLWQIFIETNLQSLENIAFN
VKKGCFTGVDGELPVAVVNDKVFVRYGDVDNLVFTNKTTLPTNVAFELFAKRKMGLTPPLSILKNLGVVATYKF
LWDYEAERPFTSYTKSVCKYTDFNEDVCVCFDNSIQGSYERFTLTTNAVLFSTVVIKNLTPIKLNFGMLNGMPV
SIKSDKGVEKLVNWYXYVRKNGQFQDHYDGFYTQGRNLSDFTPRSDMEYDFLNMDMGVFINKYGLEDFNFEHVV
GDVSKTTLGGLHLLISQFRLSKMGVLKADDFVTASDTTLRCCTVTYLNELSSKVVCTYMDLLLDDFVTILKSLD
GVISKVHEVIIDNKPYRWMLWCKDNHLSTFYPQLQSAEWKCGYAMPQIYKLQXMCLEPCNLYNYGAGIKLPSGI
LNVVKYTQLCQYLNSTTMCVPHNMRVLHYGAGSDKGVAPGTTVLKRWLPPDAIIIDNDINDYVSDADFSITGDC
TVYLEDKFDLLISDMYDGRIKFCDGENVSKDGFFTYLNGVIREKLAIGGSVAIKITEYSWNKYLYELIQRFAFW
LFCTSVNTSSSEAFLIGINYLGDFIQGPFIAGNTVHANYIFWRNSTIMSLSYNSVLDLSKFECKHKATVVVTLK
SDVNDMVLSLIKSGRLLLRNSGRFGGFSNHLVSTK
>~out: 277 to 438: Frame 1 54 aa
VVLFVQNMQICIKNMLRHIIHLHRLVLTFGYHIVLMFIICGKFLLKLIYKVLKI
>~out: 457 to 618: Frame 1 54 aa
KKGVLLVLMVSYLLQLLTTKFLFAMAMLTTWFLQIKQHCLLMLLLNCLQNEKWV
>~out: 622 to 852: Frame 1 77 aa
HHHCLFSKILVLLLHINLFYGIMKLKDLLPHILRVYVNTLILMRMFVFVLTIVFRVRMSVLRLLRTLFYFLLLS
KI
>~out: 937 to 1149: Frame 1 71 aa
LIGTXMFVKMVNFKIIMMVFTLKVGIYQTLHQEVIWSMIFLTWIWVFLLINMVLRILILNMLYMVMFQKLH
>~out: 1387 to 1572: Frame 1 62 aa
IINLIGGCCGVKITTCPLFIHSCSLLNGSVVMLCHKFISFNXCVWNLVIYIIMVLVLSCLVV
>~out: 1738 to 1935: Frame 1 66 aa
SLIMISMIMLVMQILALQVIVLLFTLKISLTYLFLICMMVELNFVMVKTSLKMVFLLILMVLLEKN
>~out: 2357 to 6142: Frame 2 1262 aa
MKLFLILLLILPLVSCFSTCNSNASISMLQLGVPDNSSTIVTGLLPVHWICANQSTSSYPANGFFYIDVGKHRSA
ALHSGYYDANQYYIYLTNKIHLNAPVTLKICKFGNTSFDFLSNVSTSHDCIVNLSFTEQLGVPLGITISGETVR
HLYNATRTFYVPAAYKLTKLSVKCYFSESCVFSVVNATITVNVTTLNGRIVNYTVCDDCNGYTDNIFSVQQDGR
PNGFPFNNWFLLTNGSTLVDGVSRLYQPLRLTCLWPVPGLKSSTGFVYFNATGSDVNCNGYQHNSVADVMRYNL
LSANSVDNLKSGVIVFKTLQYDVLFYCSNSSSGVLDTTIPFGPSSQPYYCFINSTINTTHVSTFVGILPPTVRE
VVARTGQFYINGFKYFDLGFIEAVNFNVTTASATDFWTVAFATFVDVLVNVSATNIQNLLYCDSPFEKLQCEHL
FGLQDGFYSANFLDDNVLPETYVALPIYYQHTDINFTATASFGGSCYVCKPRQVNISLNGNTSVCVRTSHFSIR
IYNRVKSGSPGDSSWHIYLKSGTCPFSFSKLNNFQKFKTICFSTVEVPGSCNFPLEATWHYTSYTIVGALYVTW
EGNSITGVPYPVSGIREFSNLVLNNCTKYNIYDYVGTGIIRSSNQSLAGGITYVSNSGNLLGFKNVSTGNIFIV
PCNQPDQVAVYQQSIIGAMTAVNESRYGLQNLLQLPNFYYVSNGGNNCTTAVMIYSNFGICADGSLIPVRPRNS
DNGISAIITANLSIPSNWTTSVQVEYLQITSTPIVVDCATYVCNGNPRCKNLLKQYTSACKTIEDALRLSAHLE
NDVSSMLTFDSNAFSLANVTSFGDYNLSSVLPQRNIHSSRIAGRSALEDLLFSKVVTSGLGTVDVDYKSCTKGL
IADLACAQYYNGIMVLPGVADAERMAMYTGSLIGGMVLGGLTSAAAIPFSLALQARLNYVALQTDVLQENQKIL
ASFNKAINNIVASFSSVNDAITHTAEAIHTVTIALNKIQDVVNQQGSALNHLTSQLRHNFQAISNSIHAIYDRL
SIQADQQVDRLITGRLAALNAFVSQVLNKYTEVRGSRRLAQQKINECVKSQSNRYGFCGNGTHIFSIVNSAPDGI
LFLHTVLLPTDYKNVKAWSGICVDGIYGYVLRQPNLVLYSDNGVFRVTSRVMFQPRLPVLSDFVQIYNCNVTFV
ISRVELHTVIPDYVDVNKTLQEFAQNLPKYVKPNFDLTPFNLTYLNLSSELKQLEAKTATNQ
>~out: 2448 to 2645: Frame 3 66 aa
VFLITLQLLSQVCCQSIGFVLIRVHLVTQPTAFSILMLVNTVVPLHSIVVIMMLTSIIFISLIKYI
>~out: 2781 to 2954: Frame 3 58 aa
LYRVKLYVCIYIMQLVLFMCRPLINLLNLVLNVTLVNPVFLVLSMPPLLLMSPHLMAV
>~out: 3126 to 3296: Frame 3 57 aa
LVYGLYLVLNLQLVLFILMPLVLMLIVTAINIILLLMLCVTILTSVLILWTILRVVL
>~out: 3546 to 3806: Frame 3 87 aa
KLSILMSRLLVPQIFGRLHLLLLLMFWLMLVQLTFKTYFIAILHLKSCSVSTCSLDCKMVFILQIFLMIMFCLRL
MLHSPFIINIRT
>~out: 3810 to 3986: Frame 3 59 aa
ILLQLHLLVVLVMFVNHARLIYLLMVTLQCVLEHLIFQLGIFITALRVVHQVTLHGIFI
>~out: 4026 to 4217: Frame 3 64 aa
IIFKSLRLFVSQPSKCLVVVIFHLKPPGITLLILLLVLCMLLGLKVIPLLVYLILSLVFVSLVI
>~out: 4227 to 4376: Frame 3 50 aa
IIVPNIIFMIMLVLELYVLQTSHLLVVLHMFLTLVIYLVLKMFPLVTFLL
>~out: 5157 to 5447: Frame 3 97 aa

Fig 3. (cont.)

20/25

VAWCSEVLHQQPPYLFLWHCKHDLTMLLYKLMCFKKIRKFWLHHLIRLLIILLLLLVALMMLLHILQRLYILLLL
HLIRFRMLLINRVVLLTISLHN
>~out: 5625 to 5774: Frame 3 50 aa
HSRRLMNVSSHNLIDMVFVAMALTSFQSSTQLQMVCFFFILFCCQLITRM
>~out: 5874 to 6065: Frame 3 64 aa
LPGSCFNLVYLFCLILCKYIIVMLLLLTYLVSSYILSYLTTLMLIKHYKSLHKTYQSMLSLILT

Alignment

>gi|12175747|ref|NP_073549.1| replicase polyprotein 1ab [Human coronavirus 229E]
gi|30179827|sp|Q05002|R1AB_CVH22 Replicase polyprotein 1ab (pp1ab) (ORF1ab polyprotein) [Includes: Replicase polyprotein 1a (pp1a) (ORF1a)] [Contains: p9; p87; p195 (Papain-like proteinases 1/2) (PL1-PRO/PL2-PRO); Peptide HD2; 3C-like proteinase (3CL-PRO) (3CLp) (M-PRO) (p34); Unknown protein 1; p5; p23; p12; Growth factor-like peptide (GFL) (p16); RNA-directed RNA polymerase (RdRp) (Pol) (p100); Helicase (Hel) (p66) (p66-HEL); Unknown protein 2; p41; Unknown protein 3]
gi|12082740|gb|AAG48591.1| replicase polyprotein 1ab [Human coronavirus 229E]
Length = 6758

Score = 1332 bits (3448), Expect = 0.0
Identities = 630/789 (79%), Positives = 695/789 (88%), Gaps = 4/789 (0%)
Frame = +3

```
Query: 3    WNCNVDMYPEFSIVCRFDTRTRSVFNLEGVNGGSLYVNKHAFHTPAYDKRAFVKLKPMPF 182
            WNCNVDMYPEFSIVCRFDTRTRS  NLEGVNGGSLYVN HAFHTPAYDKRA  KLKP PF
Sbjct: 5970 WNCNVDMYPEFSIVCRFDTRTRSTLNLEGVNGGSLYVNNHAFHTPAYDKRAMAKLKPAPF 6029

Query: 183  FYFDDSDCDVVQEQVNYVPLRASSCVTRCNIGGAVCSKHANLYQKYVEAYNTFTQAGFNI 362
            FY+DD  C+VV +QVNYVPLRA++C+T+CNIGGAVCSKHANLY+ YVE+YN FTQAGFNI
Sbjct: 6030 FYYDDGSCEVVHDQVNYVPLRATNCITKCNIGGAVCSKHANLYRAYVESYNIFTQAGFNI 6089

Query: 363  WVPHSFDVYNLWQIFIETNLQSLENIAFNVVKKGCFTGVDGELPVAVVNDKVFVRYGDVD 542
            WVP +FD YNLWQ F E NLQ LENIAFNVV KG F G DGELPVA+  DKVFVR G+ D
Sbjct: 6090 WVPTTFDCYNLWQTFTEVNLQGLENIAFNVVNKGSFVGADGELPVAISGDKVFVRDGNTD 6149

Query: 543  NLVFTNKTTLPTNVAFELFAKRKMGLTPPLSILKNLGVVATYKFVLWDYEAERPFTSYTK 722
            NLVF NKT+LPTN+AFELFAKRK+GLTPPLSILKNLGVVATYKFVLWDYEAERP TS+TK
Sbjct: 6150 NLVFVNKTSLPTNIAFELFAKRKVGLTPPLSILKNLGVVATYKFVLWDYEAERPLTSFTK 6209

Query: 723  SVCKYTDFNEDVCVCFDNSIQGSYERFTLTTNAVLFSTVVIK----NLTPIKLNFGMLNG 890
            SVC YTDF EDVC C+DNSIQGSYERFTL+TNAVLFS   +K    +L  IKLNFGMLNG
Sbjct: 6210 SVCGYTDFAEDVCTCYDNSIQGSYERFTLSTNAVLFSATAVKTGGKSLPAIKLNFGMLNG 6269

Query: 891  MPVSSIKSDKGVEKLVNWYXYVRKNGQFQDHYDGFYTQGRNLSDFTPRSDMEYDFLNMDM 1070
              ++++KS+ G  K +NW+ YVRK+G+  DHYDGFYTQGRNL DF PRS ME DFLNMD+
Sbjct: 6270 NAIATVKSEDGNIKNINWFVYVRKDGKPVDHYDGFYTQGRNLQDFLPRSTMEEDFLNMDI 6329

Query: 1071 GVFINKYGLEDFNFEHVVYGDVSKTTLGGLHLLISQFRLSKMGVLKADDFVTASDTTLRC 1250
            GVFI KYGLEDFNFEHVVYGDVSKTTLGGLHLLISQ RLSKMG+LKA++FV ASD TL+C
Sbjct: 6330 GVFIQKYGLEDFNFEHVVYGDVSKTTLGGLHLLISQVRLSKMGILKAEEFVAASDITLKC 6389

Query: 1251 CTVTYLNELSSKVVCTYMDLLLDDFVTILKSLDLGVISKVHEVIIDNKPYRWMLWCKDNH 1430
            CTVTYLN+ SSK VCTYMDLLLDDFV++LKSLDL V+SKVHEVIIDNKP+RWMLWCKDN
Sbjct: 6390 CTVTYLNDPSSKTVCTYMDLLLDDFVSVLKSLDLTVVSKVHEVIIDNKPWRWMLWCKDNA 6449

Query: 1431 LSTFYPQLQSAEWKCGYAMPQIYKLQRMCLEPCNLYNYGAGIKLPSGIMLNVVKYTQLCQ 1610
            ++TFYPQLQSAEWKCGY+MP IYK QRMCLEPCNLYNYGAG+KLPSGIM NVVKYTQLCQ
Sbjct: 6450 VATFYPQLQSAEWKCGYSMPGIYKTQRMCLEPCNLYNYGAGLKLPSGIMFNVVKYTQLCQ 6509

Query: 1611 YLNSTTMCVPHNMRVLHYGAGSDKGVAPGTTVLKRWLPPXXXXXXXXXXXYVSDADFSIT 1790
            Y NSTT+CVPHNMRVLH GAGSD GVAPGT VLKRWLP            YVSDADFS+T
Sbjct: 6510 YFNSTTLCVPHNMRVLHLGAGSDYGVAPGTAVLKRWLPHDAIVVDNDVVDYVSDADFSVT 6569

Query: 1791 GDCATVYLEDKFDLLISDMYDGRIKFCDGENVSKDGFFTYLNGVIREKLAIGGSVAIKIT 1970
            GDCATVYLEDKFDLLISDMYDGR K  DGENVSK+GFFTY+NG I EKLAIGGS+AIK+T
Sbjct: 6570 GDCATVYLEDKFDLLISDMYDGRTKAIDGENVSKEGFFTYINGFICEKLAIGGSIAIKVT 6629

Query: 1971 EYSWNKYLYELIQRFAFWTLFCTSVNTSSSEAFLIGINYLGDFIQGPFIAGNTVHANYIF 2150
            EYSWNK LYEL+QRF+FWT+FCTSVNTSSSEAF++GINYLGDF QGPFI GN +HANY+F
```

```
Sbjct: 6630 EYSWNKKLYELVQRFSFWTMFCTSVNTSSSEAFVVGINYLGDFAQGPFIDGNIIHANYVF 6689

Query: 2151 WRNSTIMSLSYNSVLDLSKFECKHKATVVVTLKDSDVNDMVLSLIKSGRLLLRNSGRFGG 2330
            WRNST+MSLSYNSVLDLSKF CKHKATVVV LKDSD+N+MVLSL++SG+LL+R +G+
Sbjct: 6690 WRNSTVMSLSYNSVLDLSKFNCKHKATVVVQLKDSDINEMVLSLVRSGKLLVRGNGKCLS 6749

Query: 2331 FSNHLVSTK 2357
            FSNHLVSTK
Sbjct: 6750 FSNHLVSTK 6758
```

>gi|13604332|gb|AAK32188.1| spike glycoprotein [Human coronavirus 229E]
Length = 1173

Score = 1391 bits (3600), Expect = 0.0
Identities = 682/1069 (63%), Positives = 833/1069 (77%), Gaps = 7/1069 (0%)
Frame = +2

```
Query: 2948 GRIVNYTVCDDCNGYTDNIFSVQQDGRIPNGFPFNNWFLLTNGSTLVDGVSRLYQPLRLT 3127
            G    +Y+VC+ C GY++N+F+V+   G IP+  F FNNWFLLTN S++VDGV R +QPL L
Sbjct: 21   GLNTSYSVCNGCVGYSENVFAVESGGYIPSDFAFNNWFLLTNTSSVVDGVVRSFQPLLLN 80

Query: 3128 CLWPVPGLKSSTGFVYFNATGSDVNCNGYQHNSVADVMRYNLNLSANSVDNLKSGVIVFK 3307
            CLW V GL+ +TGFVYFN TG    +C G+   + ++DV+RYNLN        +NL+ G I+FK
Sbjct: 81   CLWSVSGLRFTTGFVYFNGTGRG-DCKGFSSDVLSDVIRYNLNFE----ENLRRGTILFK 135

Query: 3308 TLQYDVLFYCSNSSSGVLDTTIPFGPSSQPYYCFINSTINTTHVSTFVGILPPTVREIVV 3487
            T     V+FYC+N++     D  IPFG       +YCF+N+TI        S FVG LP TVRE V+
Sbjct: 136  TSYGVVVFYCTNNTLVSGDAHIPFGTVLGNFYCFVNTTIGNETTSAFVGALPKTVREFVI 195

Query: 3488 ARTGQFYINGFKYFDLGFIEAVNFNVTTASATDFWTVAFATFVDVLVNVSATNIQNLLYC 3667
            +RTG FYING++YF LG +EAVNFNVTTA  TDF+TVA A++ DVLVNVS T+I N++YC
Sbjct: 196  SRTGHFYINGYRYFTLGNVEAVNFNVTTAETTDFFTVALASYADVLVNVSQTSIANIIYC 255

Query: 3668 DSPFEKLQCEHLQFGLQDGFYSANFLDDNVLPETYVALPIYYQHTDINFTAT---ASFGG 3838
            +S   +L+C+ L F + DGFYS + +      LP + V+LP+Y++HT I         S GG
Sbjct: 256  NSVINRLRCDQLSFDVPDGFYSTSPIQSVELPVSIVSLPVYHKHTFIVLYVDFKPQSGGG 315

Query: 3839 SCYVCKPRQVNISL-NGNTS---VCVRTSHFSIRYIYNRVKSGSPGDSSWHIYLKSGTCP 4006
             C+ C P  VNI+L N N +   +CV TSHF+ +Y+          G       W   + +G CP
Sbjct: 316  KCFNCYPAGVNITLANFNETKGPLCVDTSHFTTKYVAVYANVGR-----WSASINTGNCP 370

Query: 4007 FSFSKLNNFQKFKTICFSTVEVPGSCNFPLEATWHYTSYTIVGALYVTWSEGNSITGVPY 4186
            FSF K+NNF KF ++CFS  ++PG C  P+ A W Y+ Y  +G+LYV+WS+G+ ITGVP
Sbjct: 371  FSFGKVNNFVKFGSVCFSLKDIPGGCAMPIVANWAYSKYYTIGSLYVSWSDGDGITGVPQ 430

Query: 4187 PVSGIREFSNLVLNNCTKYNIYDYVGTGIIRSSNQSLAGGITYVSNSGNLLGFKNVSTGN 4366
            PV G+   F N+ L+ CTKYNIYD  G G+IR SN +   GITY S SGNLLGFK+V+ G
Sbjct: 431  PVEGVSSFMNVTLDKCTKYNIYDVSGVGVIRVSNDTFLNGITYTSTSGNLLGFKDVTKGT 490

Query: 4367 IFIVTPCNQPDQVAVYQQSIIGAMTAVNESRYGLQNLLQLPNFYYVSNGGNNCTTAVMIY 4546
            I+ +TPCN PDQ+ VYQQ+++GAM + N + YG  N+++LP F+Y SNG   NCT AV+ Y
Sbjct: 491  IYSITPCNPPDQLVVYQQAVVGAMLSENFTSYGFSNVVELPKFFYASNGTYNCTDAVLTY 550

Query: 4547 SNFGICADGSLIPVRPRNSSDNGISAIITANLSIPSNWTTSVQVEYLQITSTPIVVDCAT 4726
            S+FG+CADGS+I V+PRN S + +SAI+TANLSIPSNWTTSVQVEYLQITSTPIVVDC+T
Sbjct: 551  SSFGVCADGSIIAVQPRNVSYDSVSAIVTANLSIPSNWTTSVQVEYLQITSTPIVVDCST 610

Query: 4727 YVCNGNPRCKNLLKQYTSACKTIEDALRLSAHLETNDVSSMLTFDSNAFSLANVTSFGDY 4906
            YVCNGN RC  LLKQYTSACKTIEDALR SA LE+ DVS MLTFD  AF+LANV+SFGDY
Sbjct: 611  YVCNGNVRCVELLKQYTSACKTIEDALRNSARLESADVSEMLTFDKKAFTLANVSSFGDY 670

Query: 4907 NLSSVLPQRNIHSSRIAGRSALEDLLFSKVVTSGLGTVDVDYKSCTKGLSIADLACAQYY 5086
            NLSSV+P        SR+AGRSA+ED+LFSK+VTSGLGTVD DYK+CTKGLSIADLACAQYY
Sbjct: 671  NLSSVIPSLPTSGSRVAGRSAIEDILFSKIVTSGLGTVDADYKNCTKGLSIADLACAQYY 730

Query: 5087 NGIMVLPGVADAERMAMYTGSLIGGMVLGGLTSAAAIPFSLALQARLNYVALQTDVLQEN 5266
            NGIMVLPGVADAERMAMYTGSLIGG+ LGGLTSA +IPFSLA+QARLNYVALQTDVLQEN
Sbjct: 731  NGIMVLPGVADAERMAMYTGSLIGGIALGGLTSAVSIPFSLAIQARLNYVALQTDVLQEN 790

Query: 5267 QKILAASFNKAINNIVASFSSVNDAITHTAEAIHTVTIALNKIQDVVNQQGSALNHLTSQ 5446
            QKILAASFNKA+ NIV +F+ VNDAIT T++A+ TV  ALNKIQDVVNQQG++LNHLTSQ
Sbjct: 791  QKILAASFNKAMTNIVDAFTGVNDAITQTSQALQTVATALNKIQDVVNQQGNSLNHLTSQ 850

Query: 5447 LRHNFQAISNSIHAIYDRLDSIQADQQVDRLITGRLAALNAFVSQVLNKYTEVRGSRRLA 5626
```

Fig 3 (Cont.) 22/25

```
               LR NFQAIS+SI AIYDRLD+IQADQQVDRLITGRLAALN FVS  L KYTEVR SR+LA
Sbjct: 851     LRQNFQAISSSIQAIYDRLDTIQADQQVDRLITGRLAALNVFVSHTLTKYTEVRASRQLA 910

Query: 5627    QQKINECVKSQSNRYGFCGNGTHIFSIVNSAPDGLLFLHTVLLPTDYKNVKAWSGICVDG 5806
               QQK+NECVKSQS RYGFCGNGTHIFSIVN+AP+GL+FLHTVLLPT YK+V+AWSG+CVDG
Sbjct: 911     QQKVNECVKSQSKRYGFCGNGTHIFSIVNAAPEGLVFLHTVLLPTQYKDVEAWSGLCVDG 970

Query: 5807    IYGYVLRQPNLVLYSDNGVFRVTSRVMFQPRLPVLSDFVQIYNCNVTFVNISRVELHTVI 5986
                 GYVLRQPNL LY +   +R+TSR+MF+PR+P ++DFVQI NCNVTFVNISR EL T++
Sbjct: 971     TNGYVLRQPNLALYKEGNYYRITSRIMFEPRIPTMADFVQIENCNVTFVNISRSELQTIV 1030

Query: 5987    PDYVDVNKTLQEFAQNLPKYVKPNFDLTPFNLTYLNLSSELKQLEAKTA 6133
               P+Y+DVNKTLQE +  LP Y  P+  +  +N T LNL+SE+  LE K+A
Sbjct: 1031    PEYIDVNKTLQELSYKLPNYTVPDLVVEQYNQTILNLTSEISTLENKSA 1079
```

6. Sequence F

3062 Nucleotides encoding putative 3' end of Spike, hypothetical nsp 3, Envelope protein 5B, Matrix and Nucleocapsid polypeptides

```
AGCTGATCGTTGTTGATTTgAGTTGCTTAATAGGTTTGAAAATTATATCAAATGGCCTTGGTGGGTTTGGCTCAT
TATTTCTGTTGTTTTTGTTGTATTGTTGAGTCTTCTTGTGTTTTGTTGTCTTTCTACAGGTTGTTGTGGTTGTTG
CAATTGTTTAACTTCATCAATGCGAGGCTGTTGTGATTGTGGTTCAACTAAACTTCCTTATTATGAATTTGAAAA
GGTCCACGTTCAATAATGCCTTTCGGTGGCCTATTTCAACTTACTCTTGAAAGTACTATTAATAAGAGTGTGGCT
AATCTCAAATTACCACCTCATGATGTTACTGTCTTGCGTGACAATCTTAAACCTGTTACTACACTTAGTACTATC
ACTGCTTATTTGTTAGTTAGTTTGTTTGTCACTTATTTTGCTTTATTCAAACCTCTTACTGCTAGAGGTCGTGTT
GCTTGTTTTGTTTTAAAACTATTGACACTATCTGTCTATGTGCCTTTATTGGTTCTTTTTGGTATGTATCTTGAC
AGTTTTATAATTTTTTTTCTACGCTGTTGTTTCGATTCATACATGTTGGCTATTATGCCTATCTCTATAAAAATT
TTTCATTTGTTTTGTTCAATGTTACTAAACTATGCTTCGTTTCAGGCAAGTGTTGGTATCTTGAACAATCATTTT
ATGaAAATCGTTTTGCTGCTATTTATGGTGGTGACCACTATGTCGTTTTAGGTGGTGAAACTATTACTTTTGTTT
CTTTTGATGACCTTTATGTTGCTATTAGAGGTtCTTGTGAAAAGAACCTACAACTTATGCGTAAGGTTGACTTGT
ATAATGGTGCTGTCATTTACATTTTTGCCGaAGAGCCTGTTGTTGGTATAGTTTACTCCTCTCAACTATACGAAG
ATGTTCCTTCGATTAATTGATGACAATGGCATTGTCCTCAATTCTATTTTATGGCTCCTTGTTATGATATTTTTC
TTTGTGTTGGCAATGACCTTTATTAAACTGATTCAATTGTGTTTTACTTGTCATTATTTTTTTAGTAGGACATTA
TATCAACCAGTTTATAAAATTTTTCTTGCTTACCAAGATTATATGCAAATAGCACCTGTTCCAGCTGAAGTACTA
AATGTCTAAACTAAACGATGTCTAATAGTAGTGTGCCTCTTTCAGAGGTTTATGTCCATTTACGTAACTGGAACT
TTAGTTGGAATTTAATTCTAACAGTTTTTATAGTTGTGTTGCAGTATGGGCATTATAAGTATAGCAGACTTCTTT
ATGGTTTAAAGATGTCTGTTTTATGGTGTTTATGGCCACTTGTTCTAGCTTTGTCTATTTTTGACTGTTTTGTCA
ATTTTAATGTGGACTGGGTCTTTTTTGGTTTTAGTATTCTTATGTCTATTATTACACTTTGTTTATGGGTTATGT
ATTTTGTTAATAGTTTCAGACTTTGGCGCCGTGTTAAAACTTTTTGGGCTTTTAATCCTGAAACTAATGCAATCA
TCTCTCTCCAGGTTTATGGACATAATTATTACTTACCGGTGATGGCTGCACCTACAGGTGTTACATTAACACTTC
TTAGTGGTGTACTTCTTGTTGATGGCCATAAGATTGCTACTCGTGTTCAAGTGGGTCAGTTGCCTAAATATGTAA
TAGTTGCTACACCTAGTACCACAATTGTTTGTGACCGTGTTGGTCGCTCTGTTAATGAAACAAGCCAGACTGGTT
GGGCATTCTACGTCCGTGCTAAACATGGTGATTTTTCTGGTGTTGCCTCTCAGGAGGGTGTTTTGTCAGAAAGAG
AGAAGTTGCTTCATTTAATCTAAACTAAACAAAATGGCTAGTGTAAATTGGGCCGATGACAGAGCTGCTAGGAAG
AAATTTCCTCCTCCTTCATTTTACATGCCTCTTTTGGTTAGTTCTGATAAGGCACCATATAGGGTCATTCCCAGG
AATCTTGTCCCTATTGGTAAGGGTAATAAAGATGAGCAGATTGGTTATTGGAATGTTCAAGAGCGTTGGCGTATG
CGCAGGGGGCAACGTGTTGATTTGCCTCCTAAAGTTCATTTTTATTACCTAGGTACTGGACCTCATAAGGACCTT
AAATTCAGACAACGTTCTGATGGTGTTGTTTGGGTTGCTAAGGAAGGTGCTAAAACTGTTAATACCAGTCTTGGT
AATCGCAAACGTAATCAGAAACCTTTGGAACCAAAGTTCTCTATTGCTTTGCCTCCAGAGCTCTCTGTTGTTGAG
TTTGAGGATCGCTCTAATAACTCATCTCGTGCTAGCAGTCGTTCTTCAACTCGTAACAACTCACGAGACTCTTCT
CGTAGTACTTCAAGACAACAGTCTCGCACTCGTTCTGATTCTAACCAGTCTTCTTCAGATCTTGTTGCTGCTGTT
ACTTTGGCTTTAAAGAACTTAGGTTTTGATAACCAGTCGAAGTCACCTAGTTCTTCTGGTACTTCCACTCCTAAG
AAACCTAATAAGCCTCTTTCTCAACCCAGGGCTGATAAGCCTTCTCAGTTGAAGAAACCTCGTTGGAAGCGTGTT
CCTACCAGAGAGGAAAATGTTATTCAGTGCTTTGGTCCTCGTGATTTTAATCACAATATGGGGGATTCAGATCTT
GTTCAGAATGGTGTTGATGCCAAGGGTTTTCCACAGCTTGCTGAATTGATTCCTAATCAGGCTGCGTTATTCTTT
GATAGTGAGGTTAGCACTGATGAAGTGGGTGATAATGTTCAGATTACCTACACCTACAAAATGCTTGTAGCTAAG
GATAATAAGAACCTTCCTAAGTTCATTGAGCAGATTAGTGCTTTTACTAAACCCAGTTCTATCAAAGAAATGCAG
TCACAATCATCTCATGTTGCTCAGAACACAGTACTTAATGCTTCTATTCCAGAATCTAAACCATTGGCTGATGAT
GATTCAGCCATTATAGAAATTGTCAACGAGGTTTTGCATTAAATTGTTTTGTAATTCCAGTTGAATGTTTATTAT
TATTAGTTGCAACNCCCATGGTTTAGCGCATGATAAGGGTTTAGTCTACAAACGATCAAGCT
```

Hypothesised ORFs

```
>~out: 17 to 238: Frame 2          74 aa
FELLNRFENYIKWPWWVWLIISVVFVVLLSLLVFCCLSTGCCGCCNCLTSSMRGCCDCGSTKLPYYEFEKVHVQ
>~out: 223 to 723: Frame 1          167 aa
```

Fig 3 (Cont.) 23/25

```
KGPRSIMPFGGLFQLTLESTINKSVANLKLPPHDVTVLRDNLKPVTTLSTITAYLLVSLFVTYFALFKPLTAR
VACFVLKLLTLSVYVPLLVLFGMYLDSFIIFFLRCCFDSYMLAIMPISKIFHLFCSMLLNYASFQASVGILN
FMKIVLLLFMVVTTMSF
>~out: 525 to 917: Frame 3        131 aa
QFYNFFSTLLFRFIHVGYYAYLYKNFSFVLFNVTKLCFVSGKCWYLEQSFYENRFAAIYGGDHYVVLGGETIT
SFDDLYVAIRGSCEKNLQLMRKVDLYNGAVIYIFAEEPVVGIVYSSQLYEDVPSIN
>~out: 877 to 1131: Frame 1        85 aa
FTPLNYTKMFLRLIDDNGIVLNSILWLLVMIFFFVLAMTFIKLIQLCFTCHYFFSRTLYQPVYKIFLAYQDYM
APVPAEVLNV
>~out: 1140 to 1820: Frame 3        227 aa
TMSNSSVPLSEVYVHLRNWNFSWNLILTVFIVVLQYGHYKYSRLLYGLKMSVLWCLWPLVLALSIFDCFVNFN
WVFFGFSILMSIITLCLWVMYFVNSFRLWRRVKTFWAFNPETNAIISLQVYGHNYYLPVMAAPTGVTLTLLSG
LVDGHKIATRVQVGQLPKYVIVATPSTTIVCDRVGRSVNETSQTGWAFYVRAKHGDFSGVASQEGVLSEREKL
LI
>~out: 1324 to 1539: Frame 1        72 aa
LCLFLTVLSILMWTGSFLVLVFLCLLLHFVYGLCILLIVSDFGAVLKLFGLLILKLMQSSLSRFMDIIITYR
>~out: 1654 to 1815: Frame 1        54 aa
LLHLVPQLFVTVLVALLMKQARLVGHSTSVLNMVIFLVLPLRRVFCQKERSCFI
>~out: 1819 to 2964: Frame 1        382 aa
SKLNKMASVNWADDRAARKKFPPPSFYMPLLVSSDKAPYRVIPRNLVPIGKGNKDEQIGYWNVQERWRMRRGQR
DLPPKVHFYYLGTGPHKDLKFRQRSDGVVWVAKEGAKTVNTSLGNRKRNQKPLEPKFSIALPPELSVVEFEDRS
NSSRASSRSSTRNNSRDSSRSTSRQQSRTRSDSNQSSSDLVAAVTLALKNLGFDNQSKSPSSSGTSTPKKPNK
SQPRADKPSQLKKPRWKRVPTREENVIQCFGPRDFNHNMGDSDLVQNGVDAKGFPQLAELIPNQAALFFDSEVS
DEVGDNVQITYTYKMLVAKDNKNLPKFIEQISAFTKPSSIKEMQSQSSHVAQNTVLNASIPESKPLADDDSAII
IVNEVLH
>~out: 1847 to 2074: Frame 2        76 aa
IGPMTELLGRNFLLLHFTCLFWLVLIRHHIGSFPGILSLLVRVIKMSRLVIGMFKSVGVCAGGNVLICLLKFIF
T
>~out: 2078 to 2410: Frame 2        111 aa
VLDLIRTLNSDNVLMVLFGLLRKVLKLLIPVLVIANVIRNLWNQSSLLLCLQSSLLLSLRIALITHLVLAVVLQ
VTTHETLLVVLQDNSLALVLILTSLLQILLLLLLWL
>~out: 2771 to 2938: Frame 2        56 aa
LRIIRTFLSSLSRLVLLLNPVLSKKCSHNHLMLLRTQYLMLLFQNLNHWLMMIQPL
```

Alignment

>gi|13604336|gb|AAK32190.1| spike glycoprotein [Human coronavirus 229E]
Length = 1173

Score = 50.4 bits (119), Expect = 7e-06
Identities = 26/71 (36%), Positives = 31/71 (43%)
Frame = +2

```
Query: 26   LNRFENYIKWPWXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXSMRGCCDCGSTKL 205
            LNR E YIKWPW                                    S+RGCC+   STKL
Sbjct: 1105 LNRVETYIKWPWWVWLCISVVLIFVVSMLLLCCCSTGCCGFFSCFASSIRGCCE--STKL 1162

Query: 206  PYYEFEKVHVQ 238
            PYY+ EK+H+Q
Sbjct: 1163 PYYDVEKIHIQ 1173
```

>gi|12175749|ref|NP_073552.1| 4a protein [Human coronavirus 229E]
gi|138983|sp|P19739|VN4A_CVH22 Nonstructural protein 4a (ORF4a)
gi|74871|pir||MNIHHC nonstructural protein 4 - human coronavirus (strain 229E)
gi|58923|emb|CAA33682.1| unnamed protein product [Human coronavirus 229E]
gi|12082742|gb|AAG48593.1| 4a protein [Human coronavirus 229E]
Length = 133

Score = 71.6 bits (174), Expect(2) = 1e-17
Identities = 41/95 (43%), Positives = 56/95 (58%)
Frame = +1

```
Query: 253 GLFQLTLESTINKSVANLKLPPHDVTVLRDNLKPVTTLSTITAYLLVSLFVTYFALFKPL 432
           GLF L L S +N+S++N K+          +  ++K  T      + AY L+SLFV YFALFK
Sbjct: 4   GLFTLQLVSAVNQSLSNAKVSAEVSRQVIQDVKDGTVTFNLLAYTLMSLFVVYFALFKAR 63

Query: 433 TARGRVACFVLKLLTLSVYVPLLVLFGMYLDSFII 537
           + RGR A  V K+L L VYVPLL       Y+ + +I
```

```
Sbjct: 64   SHRGRAALIVFKILILFVYVPLLYWSQAYIYATLI 98
```

Score = 40.4 bits (93), Expect(2) = 1e-17
Identities = 15/30 (50%), Positives = 22/30 (73%)
Frame = +3

```
Query: 549 LLFRFIHVGYYAYLYKNFSFVLFNVTKLCF 638
            LL RF H   ++ +LYK + F++FNVT LC+
Sbjct: 102 LLGRFFHTAWHCWLYKTWDFIVFNVTTLCY 131
```

>gi|12175750|ref|NP_073553.1| 4b protein [Human coronavirus 229E]
gi|138992|sp|P19740|VN4B_CVH22 Nonstructural protein 4b (Nonstructural protein 5A) (ORF4b)
gi|74872|pir||MNIHH2 nonstructural protein 5A - human coronavirus (strain 229E)
gi|58924|emb|CAA33683.1| unnamed protein product [Human coronavirus 229E]
gi|12082743|gb|AAG48594.1| 4b protein [Human coronavirus 229E]
Length = 88

Score = 86.7 bits (213), Expect = 2e-16
Identities = 38/80 (47%), Positives = 54/80 (67%)
Frame = +1

```
Query: 640 VSGKCWYLEQSFYENRFAAIYGGDHYVVLGGETITFVSFDDLYVAIRGSCEKNLQLMRKV 819
            + GKCW+LE      +  F     YGGD ++ +G    +++ S +DLYVA+RG   +K+L L RKV
Sbjct: 1   MQGKCWFLENKALKP-FVCFYGGDQFLYIGDRIVSYFSTNDLYVALRGRIDKDLSLSRKV 59

Query: 820 DLYNGAVIYIFAEEPVVGIV 879
            +LYNG   +Y+F E P VGIV
Sbjct: 60  ELYNGECVYLFCEHPAVGIV 79
```

>gi|12175751|ref|NP_073554.1| envelope protein [Human coronavirus 229E]
gi|138994|sp|P19741|VEMP_CVH22 Envelope protein (Protein 5B)
gi|74873|pir||MNIHH3 nonstructural protein 5B - human coronavirus (strain 229E)
gi|58925|emb|CAA33684.1| unnamed protein product [Human coronavirus 229E]
gi|12082744|gb|AAG48595.1| envelope protein [Human coronavirus 229E]
Length = 77

Score = 87.8 bits (216), Expect = 3e-17
Identities = 36/76 (47%), Positives = 55/76 (72%)
Frame = +3

```
Query: 901  MFLRLIDDNGIVLNSILWLLVMIFFFVLAMTFIKLIQLCFTCHYFFSRTLYQPVYKIFLA 1080
             MFL+L+DD+ +V+N +LW +V+I    ++  +T IKLI+LCFTCH F +RT+Y P+   ++
Sbjct: 1    MFLKLVDDHALVVNVLLWCVVLIVILLVCITIIKLIKLCFTCHMFCNRTVYGPIKNVYHI 60

Query: 1081 YQDYMQIAPVPAEVLN
             YQ YM I P P  V++
Sbjct: 61   YQSYMHIDPFPKRVID 76
```

>gi|74837|pir||MMIHHC E1 membrane glycoprotein - human coronavirus (strain 229E)
gi|329573|gb|AAA45461.1| membrane protein [Human coronavirus 229E]
Length = 225

Score = 275 bits (703), Expect = 4e-72
Identities = 128/224 (57%), Positives = 159/224 (70%)
Frame = +3

```
Query: 1143 MSNSSVPLSEVYVHLRNWNFSWNLILTVFIVVLQYGHYKYSRLLYGLKMSVLWCLWPLVL 1322
             MSN +     ++  HL+NWNF WN+ILT+FIV+LQ+GHYKYSRLLYGLKM VLW LWPLVL
Sbjct: 1    MSNDNCT-GDIVTHLKNWNFGWNVILTIFIVILQFGHYKYSRLLYGLKMLVLWLLWPLVL 59

Query: 1323 ALSIFDCFVNFNVDWVFFGFSILMSIITLCLWVMYFVNSFRLWRRVKTFWAFNPETNAII 1502
             ALSIFD + N++ +W F  FS+LM++ TL +WVMYF NSFRL+RR +TFWA+NPE NAI
Sbjct: 60   ALSIFDTWANWDSNWAFVAFSLLMAVSTLVMWVMYFANSFRLFRRARTFWAWNPEVNAIT 119

Query: 1503 SLQVYGHNYYLPVMAAPXXXXXXXXXXXXXXXXXHKIATRVQVGQLPKYVIVATPSTTIVC 1682
                 V G  YY P+  AP                  H++A+ VQV  LP+Y+ VA PSTTI+
Sbjct: 120  VTTVLGQTYYQPIQQAPTGITVTLLSGVLYVDGHRLASGVQVHNLPEYMTVAVPSTTIIY 179
```

```
Query: 1683  DRVGRSVNETSQTGWAFYVRAKHGDFSGVASQEGVLSEREKLLH 1814
              RVGRSVN   + TGW FYVR KHGDFS V+S     ++E E+LLH
Sbjct: 180   SRVGRSVNSQNSTGWVFYVRVKHGDFSAVSSPMSNMTENERLLH 223
```

>gi|12175753|ref|NP_073556.1| nucleocapsid protein [Human coronavirus 229E]
gi|29840828|sp|P15130|NCAP_CVH22 Nucleocapsid protein (N structural protein) (NC)
gi|77063|pir||S08031 nucleocapsid protein - human coronavirus
gi|58933|emb|CAA35708.1| unnamed protein product [Human coronavirus 229E]
gi|12082746|gb|AAG48597.1| nucleocapsid protein [Human coronavirus 229E]
Length = 389

Score = 267 bits (682), Expect = 1e-69
Identities = 159/406 (39%), Positives = 222/406 (54%), Gaps = 31/406 (7%)
Frame = +1

```
Query: 1834  MASVNWAD---DRAARKKFPPPSFYMPLLVSSDKAPYRVIPRNLVPIGKGNKDEQIGYWN 2004
             MA+V WAD     +   R+    P S Y PLLV S++  P++VIPRNLVPI K +K++ IGYWN
Sbjct: 1     MATVKWADASEPQRGRQGRIPYSLYSPLLVDSEQ-PWKVIPRNLVPINKKDKNKLIGYWN 59

Query: 2005  VQERWRMRRGQRVDLPPKVHFYYLGTGPHKDLKFRQRSDGVVWVAKEGAKTVNTSLGNRK 2184
             VQ+R+R R+G+RVDL PK+HFYYLGTGPHKD KFR+R +GVVWVA +GAKT   T  G R+
Sbjct: 60    VQKRFRTRKGKRVDLSPKLHFYYLGTGPHKDAKFRERVEGVVWVAVDGAKTEPTGYGVRR 119

Query: 2185  RNQKPLEPKFSIALPPELSVVEFEDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 2364
             +N +P   P F+   LP   ++VVE   D
Sbjct: 120   KNSEPEIPHFNQKLPNGVTVVEEPD-----SRAPSRSQSRSQSRGRGESKPQSRNPSSDR 174

Query: 2365  XXXXXXXLVAAVTLALKNLGFDN----------------QXXXXXXXXXXXXXXXXXXXLS 2496
                       ++ AV   ALK+LGFD                                     S
Sbjct: 175   NHNSQDDIMKAVAAALKSLGFDKPQEKDKKSAKTGTPKPSRNQSPASSQTSAKSLARSQS 234

Query: 2497  QPRADKPSQLKKPRWKRVPTRE--ENVIQCFGPRDFNHNMGDSDLVQNGVDAKGFPQLAE 2670
                 ++   +++KPRWKR P   +     NV QCFGPRD +HN G + +V NGV AKG+PQ AE
Sbjct: 235   SETKEQKHEMQKPRWKRQPNDDVTSNVTQCFGPRDLDHNFGSAGVVANGVKAKGYPQFAE 294

Query: 2671  LIPNQAALFFDSEVSTDEVGDNVQITYTYKMLVAKDNKNLPKFIEQISAFTKPSSIKEMQ 2850
             L+P+ AA+ FDS + + E G+ V +T+T ++ V KD+ +L KF+E+++AFT       +EMQ
Sbjct: 295   LVPSTAAMLFDSHIVSKESGNTVVLTFTTRVTVPKDHPHLGKFLEELNAFT-----REMQ 349

Query: 2851  SQSSHVAQNTVLNASIPE----------SKPLADDDSAIIEIVNEV 2958
                     Q+ +LN S  E          ++P+ D+ S    +I++EV
Sbjct: 350   -------QHPLLNPSALEFNPSQTSPATAEPVRDEVSIETDIIDEV 388
```

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/NL04/000805

International filing date: 18 November 2004 (18.11.2004)

Document type: Certified copy of priority document

Document details: Country/Office: EP
Number: 03078613.1
Filing date: 18 November 2003 (18.11.2003)

Date of receipt at the International Bureau: 31 January 2005 (31.01.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)